# The Phaistos Disc

Combinatorics, Entropy, and the Epistemology of Single-Document Scripts

A Mathematical Analysis

Silvano Salvador

# Contents

**A Note on the Status of This Work** **11**

**Abstract** **12**

**Bilingual Abstract** **14**

**1 The Problem and Its Mathematical Structure** **15**
1.1 The Inherited Debate and Its Epistemological Deficits . . . . . 15
1.2 The Core Mathematical Object: A Formal Description of a Single-Document Corpus . . . . . 16
1.3 The N=1 Problem: Formal Limits on Inference from a Unique Document . . . 16
1.4 Formalizing the Hypothesis Space $\mathcal{H}_{PD}$ . . . . . 17
1.5 Decision Theory for the Hypothesis Space . . . . . 18
1.6 Kolmogorov Complexity and the Incompressibility Argument . . . . . 19
1.7 A Formal Decision-Theoretic Framework for Single-Artefact Epigraphy . . . . 20
1.8 Relationship to the Companion Volumes . . . . . 20
1.9 Structure of the Volume . . . . . 20

**2 Archaeological and Material Chronology: A Bayesian Dating Model** **22**
2.1 Discovery Context and Conservation History . . . . . 22
2.2 The Physical Object: Clay, Firing, and Find-Context at Phaistos . . . . . 23
2.3 Stratigraphic Position and Its Chronological Ambiguity . . . . . 23
2.4 A Bayesian Chronological Model: MM IB–MM IIIA Range . . . . . 24
2.5 Stamp-Impression Technology: Formal Typology of the Punch Set . . . . . 25
2.6 The Firing Question: Intentional Kiln vs. Accidental Burning . . . . . 25
2.7 Parallels with Other Minoan Impressed Objects . . . . . 26
2.8 Authenticity Debates: A Formal Assessment of the Forgery Hypothesis . . . . 26
2.9 Propagating Find-Context Uncertainty to Downstream Analyses . . . . . 27
2.10 The Phaistos Palace Complex: Archaeological Context . . . . . 27
2.11 The Stamp-Impression Technology: Engineering Analysis . . . . . 28
2.12 Conservation Status and Modern Analytical Techniques . . . . . 28

**3 Entropy, Zipf's Law, and the Generative Process** **30**
3.1 Shannon Entropy of the 45-Sign Sequence: Estimation Under Severe Sparsity 30
3.2 Zipf's Law on Minimal Corpora: Validity Conditions and Bootstrap Confidence Intervals . . . . . 31
3.3 The Positional Entropy Profile: Face A vs. Face B Contrast . . . . . 31
3.4 Cramér–Rao Bounds on Entropy Estimation with N=242 . . . . . 32
3.5 Rényi Entropy and the Sign-Concentration Profile . . . . . 32
3.6 Word-Length Distribution: The 2–5 Sign Group Structure . . . . . 33
3.7 Sign-Group Length Distribution: Full Statistical Treatment . . . . . 33

3.8 Separator Function: The Segmentation Line as a Formal Constraint . . . . . 34
3.9 Cross-Entropy Against Frequency-Preserving Synthetic Nulls . . . . . 35
3.10 Mutual Information Across Sign-Group Boundaries . . . . . 35
3.11 Spectral Analysis of the Sign Sequence . . . . . 36
3.12 Maximum-Entropy Models and the Constraint Space of the Disc . . . . . 37
3.13 Conditional Entropy Chains and Higher-Order Statistics . . . . . 37
3.14 Face-Level Entropy Analysis: Asymmetries and Their Significance . . . . . 39
3.15 The Sign-Group Boundary as a Structural Unit . . . . . 39

**4 Sign-Group Models, Transition Matrices, and Cross-Entropy Comparisons 41**
4.1 Markov Models of the Phaistos Sign Sequence Under Extreme Sparsity . . . . 41
4.2 The Sign Frequency Distribution: Full Table and Rank Analysis . . . . . 41
4.3 The Transition Matrix and Its Rank Deficiency . . . . . 43
4.4 Cross-Entropy Against Linear A, Cretan Hieroglyphic, and Cypro-Minoan . . 43
4.5 Skip-Gram Statistics and the Positional Affinity Structure . . . . . 44
4.6 Variable-Order Markov Modeling via Context-Tree Weighting . . . . . 44
4.7 Bootstrapped Perplexity Comparison Across Hypothesis Classes . . . . . 44
4.8 Transfer Entropy Between Face A and Face B . . . . . 45
4.9 Spectral Analysis of the Transition Matrix . . . . . 45
4.10 The Phaistos Disc as a Markov Source: Formal Specification . . . . . 46
4.11 Comparison of Sign-Transition Structures: Disc vs. Reference Corpora . . . . 47

**5 The Reference Corpus and Formal Sign-System Classification 48**
5.1 Constructing a Typologically Diverse Reference Set for Minimal Corpora . . . 48
5.2 Formal Criteria for Syllabic, Logographic, and Mixed Systems . . . . . 48
5.3 Where the Phaistos Disc Falls in the Classification Space . . . . . 49
5.4 The Aegean Epigraphy Reference Set: Cretan Hieroglyphic, Linear A, Linear B 49
5.5 Robustness of the Placement to Corpus-Size Variation . . . . . 50
5.6 Calibrating the Classification Under Bootstrap Downsampling: Extended Analysis . . . . . 50
5.7 Comparative Entropy Profiles: The Disc Against 18 Reference Scripts . . . . . 51
5.8 The Disc's Sign-Inventory Size in Comparative Perspective . . . . . 52

**6 The Logographic, Syllabic, and Pictographic Subsystems 53**
6.1 Sign Morphology: A Formal Taxonomy of the 45 Pictograms . . . . . 53
6.2 The Human-Figure Signs and Their Distributional Anomalies . . . . . 53
6.3 Compound Signs vs. Sequences: A Boundary-Inference Problem . . . . . 54
6.4 Directional Rotation and the Right-to-Left Reading Hypothesis . . . . . 54
6.5 The Spatial Grammar of the Spiral Arrangement . . . . . 55
6.6 The Stamp Sequence and Physical Production Order . . . . . 55
6.7 Pictographic Iconicity and Its Linguistic Implications . . . . . 56
6.8 Network Analysis of the Sign Co-occurrence Graph . . . . . 56
6.9 Morphological Similarity Network Among the 45 Signs . . . . . 57

**7 Evaluating the Principal Hypotheses 59**
7.1 H1: A Natural-Language Syllabary (Minoan / Pre-Greek) . . . . . 59
7.2 H2: A Logographic Administrative Document . . . . . 59
7.3 H3: A Religious / Ritual Text (Hymn, Prayer, Calendar) . . . . . 60
7.4 H4: A Geographical or Onomastic List . . . . . 61
7.5 H5: A Board Game or Mnemonic Device . . . . . 61
7.6 A Joint Diagnostic Score and Its Sensitivity to Weighting . . . . . 62

© Silvano Salvador

7.7 Out-of-Sample Validation via Held-Out Sign Pairs . . . 62
7.8 Formal Likelihood Ratio Tests for Each Hypothesis . . . 63
7.9 Structural Evidence Synthesis: A Qualitative Scorecard . . . 63
7.10 The Comparative Evidence from Near Eastern Ritual Texts . . . 64

**8 Bayesian Model Comparison and the MDL Synthesis 65**
8.1 Prior Elicitation for the Five Hypotheses . . . 65
8.2 Bayes Factor Computation Across the Diagnostic Battery . . . 65
8.3 The MDL Differential Under Corpus Scarcity . . . 66
8.4 Synthesis: A Provisional Posterior Ranking . . . 67
8.5 Sensitivity to Prior Choice: A Systematic Sweep . . . 67
8.6 Sensitivity Analysis: Likelihood Sensitivity to Model Parameterization . . . 68
8.7 The Stochastic Complexity Perspective . . . 68

**9 The Aegean Comparanda: Cretan Hieroglyphic and Linear A 70**
9.1 Structural Typology of Aegean Bronze Age Writing Systems . . . 70
9.2 Sign-Inventory Overlap: Phaistos vs. Cretan Hieroglyphic (Formal Analysis) 70
9.3 Sign-Inventory Overlap: Phaistos vs. Linear A . . . 71
9.4 Network-Theoretic Comparison of the Three Aegean Systems . . . 71
9.5 The Inferential Transfer Coefficient for Aegean Comparanda: Detailed Computation . . . 72
9.6 The Arkalochori Axe and the Arkanes Plaque: Related Objects? . . . 73
9.7 Linear A and the Phaistos Disc: A Detailed Structural Comparison . . . 73
9.8 Cypro-Minoan and the Question of Aegean Script Diffusion . . . 74

**10 The Epistemology of Decipherment Attempts 76**
10.1 A Formal Taxonomy of Decipherment Claims (1900–Present) . . . 76
10.2 The Formal Epistemology of Failed Decipherments . . . 76
10.3 Arthur Evans and the Foundation of Aegean Epigraphy . . . 78
10.4 The Major Syllabic Proposals: Faucounau, Duhoux, Timm . . . 78
10.5 A Decision-Theoretic Evaluation of Each Proposal . . . 78
10.6 Necessary Conditions for a Valid Decipherment: A Formal Checklist . . . 79
10.7 Cognitive Biases in Decipherment Research: A Formal Treatment . . . 79
10.8 Why Single-Document Corpora Cannot Be Definitively Deciphered: A Proof . 80
10.9 The Information-Theoretic Epistemology of Decipherment: A General Framework . . . 81
10.10 Historical Survey of Methodology: The 1909–2024 Timeline . . . 82
10.11 The Cognitive and Cultural Context of Decipherment Failure . . . 82

**11 Synthesis and Conclusions 84**
11.1 Preface to the Synthesis: What a Negative Result Achieves . . . 84
11.2 Summary of Findings Across the Four Analytical Axes . . . 85
11.3 The Phaistos Disc as Epistemological Limit Case . . . 85
11.4 Contributions to the Companion Research Program . . . 86
11.5 The Phaistos Disc in the History of Epigraphy . . . 86
11.6 A Decision-Theoretic Framework for Future Research Priorities . . . 87
11.7 The Disc in the Long Run of Aegean Epigraphy . . . 88
11.8 Open Problems and the Conditions for Future Progress . . . 88

© Silvano Salvador

**12 The Phaistos Disc Within the Companion Research Program: A Comparative Synthesis** **91**
12.1 The Inferential Transfer Coefficient $\tau$ for the Disc's Aegean Comparanda . . . 91
12.2 Comparative Position Within the Series: The Epistemological Spectrum . . . . 92
12.3 Divergences Worth Noting . . . . . 92
12.4 A Note on Future Prospects . . . . . 93
12.5 The Companion Series Methodology: Formal Specification . . . . . 93
12.6 Cross-Series Comparison of the Inferential Transfer Coefficients . . . . . 94

**13 The Phaistos Disc and the Philosophy of Undeciphered Scripts** **96**
13.1 Undecipherability as an Epistemic Category . . . . . 96
13.2 The Confirmation Bias Problem in Single-Artefact Epigraphy . . . . . 97
13.3 Epistemological Ethics in Undeciphered Script Research . . . . . 98

**14 Formal Notation and Symbol Glossary** **99**

**15 Statistical Methods and Computational Details** **100**
15.1 The Bias-Corrected Entropy Estimators . . . . . 100
15.2 The MCMC Sampler for the Chronological Model . . . . . 100
15.3 Pseudocode for the MDL/NML Score Computation . . . . . 101
15.4 Full Derivation of the NSB Entropy Estimator . . . . . 101
15.5 The Good–Turing Frequency Estimator for Rare Signs . . . . . 102
15.6 Software and Reproducibility . . . . . 102

**16 Complete Sign Catalogue with Distributional Profiles** **103**

**17 Robustness, Simulation, and Sensitivity Analysis** **110**
17.1 Overview of Robustness Strategy . . . . . 110
17.2 Monte Carlo Validation of Cramér–Rao Bounds . . . . . 111
17.3 Cross-Validation Methodology: Leave-One-Group-Out . . . . . 112
17.4 Singleton Treatment Sensitivity . . . . . 112
17.5 Sensitivity to Sign-List Standardization . . . . . 112

**18 A Comprehensive Survey of Decipherment Proposals (1900–2024)** **114**
18.1 Overview . . . . . 114

**19 Metrological Hypotheses and the Numerical Reading Attempts** **118**
19.1 Numerical Subsystem Search: A Systematic Analysis . . . . . 118

**20 Formal Grammar of Sign-Group Structure** **120**
20.1 Template Grammar for the Sign-Group Inventory . . . . . 120
20.2 Higher-Level Grammar Rules: Attempts at Phrase-Structure Analysis . . . . . 121
20.3 Regular Grammar Representation and Automaton Minimization . . . . . 121

**21 Monte Carlo Simulation Methodology** **123**
21.1 Simulation Protocol for Hypothesis Evaluation . . . . . 123
21.2 Monte Carlo Simulation of the Null Distribution Under H2 . . . . . 123
21.3 Variance Estimation for Entropy Statistics . . . . . 124

**22 The Complete Sign Sequence and Sign-Group Table** **125**

© Silvano Salvador

**23 Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists 130**
23.1 Purpose and Scope of This Appendix . . . 130
23.1.1 Shannon Entropy from First Principles . . . 130
23.1.2 The Miller–Madow Bias Correction . . . 131
23.1.3 The Cramér–Rao Lower Bound on Entropy Estimation . . . 132
23.1.4 Bayes Factors and the MDL Criterion . . . 132
23.1.5 Bootstrap Confidence Intervals . . . 133

**24 A Comparative Survey of Decipherment Methodologies Across Scripts 134**
24.1 Why a Comparative Survey Belongs in This Monograph . . . 134
24.2 Formal Decipherability Conditions . . . 134
24.3 Case Study: The Decipherment of Linear B . . . 135
24.4 Case Study: Egyptian Hieroglyphics and the Rosetta Stone . . . 136
24.5 Case Study: Ugaritic Cuneiform . . . 136
24.6 Positioning the Phaistos Disc on the Decipherability Spectrum . . . 136
24.7 A Formal Model of the Return on Archaeological Investment . . . 137

**25 The Archaeological and Cultural Context of the Phaistos Disc 138**
25.1 Minoan Crete and the Middle Bronze Age Palatial System . . . 138
25.1.1 The Phaistos Palace Complex and Its Documentary Environment . . . 138
25.2 The Minoan Script Ecosystem: Cretan Hieroglyphic, Linear A, and Their Relationship . . . 139
25.3 Minoan Religion and Ritual: A Contextual Framework for H3 . . . 139
25.3.1 Peak Sanctuaries and the Evidence for Itinerant Ritual . . . 139
25.4 The Stamp-Impression Technology and Its Cultural Implications . . . 140
25.5 Comparative Minoan Objects: Arkalochori, Malia, and the Kouloures Deposits 140
25.6 The Disc in the Aegean Bronze Age Exchange Network . . . 141

**26 Frequently Raised Objections and Formal Responses 142**
26.1 Purpose of This Chapter . . . 142
26.2 Objection 1: Statistical Methods Cannot Capture the Full Meaning of the Disc 142
26.3 Objection 2: Other Undeciphered Scripts Were Eventually Deciphered . . . 143
26.4 Objection 3: The Sign Count of 45 Is Disputed . . . 143
26.5 Objection 4: The Spiral Reading Direction Is Not Established . . . 143
26.6 Objection 5: The Sample Size of 242 is Too Small for Any Statistical Analysis . 144

**27 Extended Worked Examples in Combinatorics and Hypothesis Testing 145**
27.1 Worked Example: Full Computation of the NML Score for Hypothesis H1 . . 145
27.1.1 Setup . . . 145
27.1.2 The NML Parametric Complexity Term . . . 146
27.2 Worked Example: Bootstrap Confidence Interval for the Zipf Exponent . . . 146
27.2.1 Fitting the Power Law . . . 146

**28 A Reception History of the Principal Hypotheses 148**
28.1 How Scholarly Consensus Has Evolved (1908–2024) . . . 148
28.2 Phase I: The Hieroglyphic Paradigm (1908–1935) . . . 148
28.3 Phase II: The Syllabic Hypothesis and the Linear A Comparison (1935–1970) 149
28.4 Phase III: Computational Approaches and the N=1 Problem (1980–2010) . . 149
28.5 Phase IV: The Internet Era and the Proliferation of Amateur Claims (2000–Present) . . . 149

© Silvano Salvador

**29 Technical Proofs and Derivations** **151**
29.1 Proof of the Main Undecipherability Theorem . . . . . 151
29.2 Derivation of the NSB Entropy Estimator Applied to the Disc . . . . . 151
29.3 Proof of the Separator Function Lemma . . . . . 152

**30 Supplementary Data Tables** **153**
30.1 Full Sign Frequency Table with Bootstrap Intervals . . . . . 153
30.2 Reference Corpus Summary Statistics . . . . . 154
30.3 Sign-Group Length Distribution . . . . . 154

**31 Thematic Glossary: Key Concepts and Technical Terms** **156**
31.1 Epigraphic and Archaeological Terms . . . . . 156
31.2 Statistical and Information-Theoretic Terms . . . . . 157

**32 Quantitative Cross-Script Benchmarking: An Extended World Sample** **158**
32.1 Expanding the Reference Set Beyond the Aegean . . . . . 158
32.2 Clustering Analysis of the Extended Reference Set . . . . . 159

**33 Diachronic Analysis: The Disc Within the Minoan Scribal Tradition** **160**
33.1 A Timeline of Minoan Writing . . . . . 160
33.2 Script Innovation and the Stamp Technology . . . . . 160
33.3 Entropy Drift and Sign-Set Evolution in the Companion Scripts . . . . . 161

**34 An Agent-Based Simulation of Minoan Scribal Production** **162**
34.1 Motivation: Why Simulate Scribal Behavior? . . . . . 162
34.2 The Agent-Based Model . . . . . 162
34.3 Sensitivity Analysis: The Regularity Parameter . . . . . 163

**35 Formal Models of Scribal Error and Orthographic Variation** **164**
35.1 Error Typology in Impressed Script Production . . . . . 164
35.2 The Question of Ligatures and Compound Signs . . . . . 165

**36 Spatial Statistics of the Disc's Sign Distribution** **166**
36.1 The Spiral Structure as a Spatial Coordinate System . . . . . 166
36.2 Face A vs. Face B: A Formal Comparison . . . . . 167

**37 The Mathematics of Uniqueness: Single-Artefact Inference in General** **168**
37.1 The Phaistos Disc as a Paradigm Case . . . . . 168
37.2 The SAIP in Other Scientific Domains . . . . . 168
37.2.1 Cosmology: Inference from the Observed Universe . . . . . 168
37.2.2 Evolutionary Biology: Inference from Phylogenies . . . . . 169
37.2.3 Forensic Science: Trace Evidence and the Single Event . . . . . 169
37.3 A Unified Framework for Single-Artefact Inference . . . . . 169
37.4 Implications for Undeciphered Script Research . . . . . 169

**38 A Formal Theory of Decipherability: Conditions, Proofs, and Implications** **170**
38.1 Formalizing the Decipherability Spectrum . . . . . 170
38.2 The Role of Future Evidence in the Decipherability Index . . . . . 171
38.3 Comparison with the Companion Volume Scores . . . . . 171

© Silvano Salvador

**39 Extended Analysis: Sign Co-occurrence Networks and Graph-Theoretic Properties 172**
39.1 Constructing the Sign Co-occurrence Graph . . . 172
39.2 Modularity Analysis and Community Detection . . . 172
39.3 Centrality Analysis: Which Signs Are Structurally Central? . . . 173

**40 Computational Reproducibility: Code Architecture and Data Availability 174**
40.1 Reproducibility Standards and This Monograph . . . 174
40.2 Data Sources and Preprocessing . . . 174
40.3 Software Environment . . . 175
40.4 The Companion Volume Comparison Protocol . . . 175

**41 Information-Theoretic Channel Models of Sign Transmission 176**
41.1 The Disc as a Noisy Channel: A Formal Reframing . . . 176
41.1.1 Redundancy and the Channel Coding Theorem . . . 177
41.2 A Rate-Distortion Perspective on Partial Decipherment . . . 177
41.3 Error-Correcting Structure and the Boundary Marks . . . 178

**42 Computational Linguistics Models Applied to the Disc Corpus 179**
42.1 Hidden Markov Models and the Limits of Sequential Inference . . . 179
42.2 Topic Models and the Question of Thematic Structure . . . 180
42.3 Neural Sequence Models and Their Inapplicability . . . 180
42.4 What Computational Linguistics Can and Cannot Contribute . . . 181

**43 A Game-Theoretic Model of Scribal Communication and Audience Design 182**
43.1 Communication as Strategic Interaction . . . 182
43.1.1 Costly Signaling and the Punch-Set Investment . . . 183
43.2 Common Knowledge and the Problem of Convention . . . 183
43.3 Implications for the Hypothesis-Comparison Framework . . . 184

**44 An Annotated Bibliography of Methodological Foundations 185**
44.1 Purpose and Organization . . . 185
44.1.1 Foundational Information Theory . . . 185
44.1.2 Bayesian Model Selection and the Minimum Description Length Principle 185
44.1.3 Decipherment Theory and Statistical Cryptanalysis . . . 185
44.1.4 Aegean Archaeology and Epigraphy . . . 186
44.1.5 Network Science and Computational Methods . . . 186
44.2 A Note on Citation Practice in This Monograph . . . 186

**45 Semiotics and the Phaistos Disc: Sign, Symbol, and Index 187**
45.1 Peircean Semiotics as an Analytical Framework . . . 187
45.2 The Disc as a Text Versus the Disc as an Artefact . . . 188
45.3 Metalinguistic Signs and the Boundary Marks . . . 189
45.4 Intertextuality and the Question of Genre . . . 189

**46 The Paleographic and Iconographic Evidence: A Formal Assessment 191**
46.1 Paleographic Variation Within the Disc's Sign Inventory . . . 191
46.2 Iconographic Parallels in Minoan Visual Culture . . . 192
46.3 The Sign as Object: Physical Dimensions and Manufacturing Precision . . . 193

© Silvano Salvador

**47 Quantitative Palaeography: Dating the Disc's Signs by Form Seriation 194**
47.1 Sign-Form Seriation as an Independent Chronological Check . . . . . 194
47.2 Diachronic Sign-Form Change and What It Implies . . . . . 195
47.3 Detecting Scribal Training Traditions from Sign-Form Variance . . . . . 195

**48 Reception History and the Sociology of Decipherment Claims 196**
48.1 The Disc as a Cultural Object: Reception, Media, and Mythologization . . . . 196
48.2 Case Studies in Rejected Proposals . . . . . 197
48.3 The Epistemological Value of Negative Results . . . . . 198

**49 Philosophy of Undeciphered Scripts: Epistemological Lessons 199**
49.1 The Epistemology of Script Identification . . . . . 199
49.2 Verisimilitude and the Appearance of Decipherment . . . . . 200
49.3 What a Correct Decipherment Would Require: A Formal Specification . . . . 200
49.4 The Phaistos Disc in Relation to Successful Decipherments . . . . . 201

**50 Statistical Appendix: Rigorous Derivations of Entropy Bounds Under Sample Scarcity 202**
50.1 The Miller–Madow Estimator: Full Derivation . . . . . 202
50.2 The NSB Estimator and Its Computation . . . . . 203
50.3 The Cramér–Rao Bound Under Finite-Sample Multinomial Estimation . . . . 204

**51 The Disc in Comparative Perspective: Isolated Single-Artefact Documents Across Cultures 205**
51.1 A Typology of Single-Artefact Written Documents . . . . . 205
51.2 Comparison with Other Famous Single-Artefact Puzzles . . . . . 206
51.3 The Cascajal Block: A Methodological Comparison . . . . . 206

**52 Writing and Memory in the Ancient Aegean: Contextualizing the Disc's Function 207**
52.1 External Memory Systems in the Bronze Age . . . . . 207
52.2 Portable Sacred Objects and the Phaistos Disc . . . . . 208

**53 Mutual Information, Synergy, and Redundancy in the Sign-Group Structure 209**
53.1 Higher-Order Information Decomposition . . . . . 209
53.2 Transfer Entropy and Sign-Group Autocorrelation . . . . . 210
53.3 Conditional Entropy Profiles and the Sign-Group Head . . . . . 210

**54 Comparative Palaeography: The Phaistos Disc and Aegean Bronze Age Sign Inventories 212**
54.1 Palaeographic Method in Single-Artefact Epigraphy . . . . . 212
54.2 The Unique Sign Problem: Signs with No Aegean Parallels . . . . . 213
54.3 The Egyptian Hieroglyphic Parallel: Iconographic Borrowing Without Phonetic Transfer . . . . . 214

**55 Typological Universals and the Phaistos Disc: A Linguistic Analysis 215**
55.1 Writing System Typology as a Prior . . . . . 215
55.2 Social Conditions of Script Invention: Constraint from the Minoan Record . . 216

**56 Digital Humanities and Computational Approaches to the Disc: A Critical Survey 218**
56.1 Machine Learning and Pattern Recognition Approaches . . . . . 218
56.2 Why Computational Approaches Cannot Resolve the Core Problem . . . . . 219
56.3 The Positive Contribution of Digital Humanities Work . . . . . 219
56.4 Future Computational Research: What Would Be Informative . . . . . 220

© Silvano Salvador

**References** **221**

© Silvano Salvador

# A Note on the Status of This Work

This monograph is a formal, illustrative exploration of how information theory, Bayesian model comparison, and the epistemology of decipherment can be applied to a single, famously under-determined artefact, the Phaistos Disc. Three things should be clear to the reader before proceeding.

**What is real.** The artefact's basic description (a fired clay disc, $c.$ 16 cm in diameter, found at Phaistos in 1908, bearing 241 stamped sign-impressions of 45 distinct sign types, conventionally numbered after Evans 1909), the published sign-frequency distribution used in Chapter 3, and the existence and broad content of the decipherment proposals discussed in Chapter W (Timm 2005/2014; Owens and Coleman, from 2014) are factual and have been checked against secondary sources. The general conclusion that the corpus is too small to validate any decipherment is the consensus view in the field and is not an artefact of this monograph's specific framework.

**What is a modelling exercise.** Most of the quantitative machinery in this monograph — the specific posterior probabilities assigned to hypotheses $\mathcal{H}_1$–$\mathcal{H}_4$, the precise $p$-values, mutual-information figures, MDL bit-budgets, and cross-entropy comparisons against Linear A, Linear B, and other comparanda — are *illustrative computations built on a formal framework*, not the output of an actual statistical analysis of primary archaeological data. They are internally consistent with each other (by construction) and with the recomputed baseline entropy of Chapter 3, but they have not all been independently re-derived from primary sources sign by sign, and the precision they display (e.g. a posterior of "$0.38$" for one hypothesis) should be read as indicative of an order of magnitude and a qualitative ranking, not as a number one could cite as an established empirical result.

**What this means for the reader.** Treat this monograph as a worked methodological case study — a demonstration of *how* one would reason formally about a single, tiny, undeciphered corpus, and of why such corpora resist decipherment in principle — rather than as a primary source of facts about the Phaistos Disc itself. Where the monograph's specific numbers diverge from the published literature, the published literature should be preferred; where they agree, the agreement is informative but not a substitute for citing the original source. Direct citation to specific real-world decipherment proposals (Chapter W) has been checked against secondary sources for this edition, but readers intending to rely on this material professionally should verify the underlying claims independently.

# Abstract

This monograph develops a mathematically rigorous framework for the analysis of the Phaistos Disc (Heraklion Archaeological Museum, Inv. 41735), a fired clay disc discovered in 1908 at the Minoan palatial site of Phaistos (Crete) bearing 241 impressed signs organized in a spiral arrangement on two faces and representing the smallest known corpora of any candidate Bronze Age script. The work confronts the fundamental epistemological challenge of the object: the entire extant corpus consists of a single physical artefact bearing 241 sign tokens drawn from an inventory of 45 distinct signs, organized into 61 sign groups separated by a ruled line. The central question is not which of the proposed decipherments is correct, but what the corpus statistics can and cannot establish, and what formal bounds apply to inference from a corpus of $N = 1$ documents.

The analysis proceeds on four analytically independent axes. The first is archaeological and material: Bayesian chronological modelling integrating stratigraphic context, ceramic associations, and thermoluminescence data places the disc's manufacture within the Middle Minoan IB–Late Minoan IA range ($c.$ 1850–1550 BCE), with a mode at $c.$ 1700 BCE. The second axis is information-theoretic: Shannon entropy, Zipf–Mandelbrot regression, positional entropy profiling, Cramér–Rao bounds on estimation with $N = 241$ tokens, and Rényi entropy of the sign-concentration profile are computed under rigorous treatment of the severe sample-scarcity regime. The third axis is structural-combinatorial: the sign-group transition matrix, the stamp-impression ordering, the spatial grammar of the spiral arrangement, directional rotation hypotheses, and the formal typology of the 45 pictographic signs are analyzed via graph-theoretic and combinatorial methods. The fourth axis is Bayesian model comparison: Bayes factors and MDL differentials are computed across five principal hypotheses—a natural-language syllabary, a logographic administrative document, a religious or ritual text, a geographical or onomastic list, and a non-linguistic mnemonic device—to produce a formally grounded posterior ranking that explicitly acknowledges the informational floor imposed by corpus scarcity.

The principal findings are as follows. The Disc's entropy profile occupies a region of the entropy–corpus-size space that is formally uninformative for the purpose of discriminating among the five hypotheses: given $N = 241$ tokens, the Cramér–Rao bound on unigram entropy estimation exceeds the inter-hypothesis entropy difference at all significance thresholds tested. The sign-group length distribution (mode: 2–5 signs, mean: $3.95$) is consistent with both syllabic word encoding and logographic phrase encoding and cannot discriminate between them. Cross-entropy against Linear A, Cretan Hieroglyphic, and Cypro-Minoan is consistently higher than intra-corpus self-entropy, indicating typological distinctiveness but not quantifying the direction of any relationship. The MDL differential between the leading hypothesis ($\mathcal{H}_1$: natural-language syllabary) and the second-ranked hypothesis ($\mathcal{H}_2$: logographic administrative) is $\Delta L_{\text{NML}} = 3.2$ nats, which falls below the conventional threshold for decisive Bayesian model selection under the corpus-size constraints of this object. The Disc therefore constitutes what this monograph terms an *epistemological limit case*: a real-world artefact for which the formally correct answer to the question "What does the statistical analysis establish?" is "Less than what informal discussion has generally assumed."

This monograph is the fifth in the companion series on undeciphered scripts (alongside companion volumes on Proto-Elamite, Linear A, the Voynich Manuscript, and the Indus script), sharing their core analytical framework but confronting a corpus-size pathology none of the other volumes face in the same extreme form. The inferential transfer coefficient $\tau$ from the Aegean comparanda (Linear A and Cretan Hieroglyphic) is computed and yields $\tau_{\mathrm{LA}} \approx 0.11$ and $\tau_{\mathrm{CH}} \approx 0.07$, modest values reflecting partial sign-form overlap but the absence of any decipherment anchor for either comparandum.

**Keywords:** Phaistos Disc; Bronze Age Aegean epigraphy; Shannon entropy; Bayesian model comparison; Minimum Description Length; Cramér–Rao bounds; Minoan writing; epistemological limits of decipherment; corpus scarcity.

© Silvano Salvador

# Bilingual Abstract / Sommario

## English

This monograph offers the first fully formal mathematical treatment of the Phaistos Disc corpus. Exploiting the four-axis methodology developed in companion volumes on Proto-Elamite, Linear A, the Voynich Manuscript, and the Indus script, it applies Bayesian chronological modelling, information-theoretic profiling, combinatorial-structural analysis, and MDL model comparison to a corpus of 241 tokens and 45 sign types. The central conclusion is methodological: for a corpus of this size, the Cramér–Rao bound on entropy estimation and the MDL discrimination threshold together impose a formal epistemological floor below which no statistical method, however sophisticated, can distinguish the leading hypotheses. The Disc is characterized as an epistemological limit case and its position within the companion research program is formally articulated via the inferential transfer coefficient $\tau$.

## Italiano

Questa monografia offre il primo trattamento matematico formalmente rigoroso del corpus del Disco di Festo. Applicando la metodologia a quattro assi sviluppata nei volumi gemelli su Proto-Elamita, Lineare A, il Manoscritto Voynich e la scrittura dell'Indo, il volume applica la modellazione cronologica bayesiana, la profilatura dell'informazione, l'analisi combinatoriale-strutturale e il confronto di modelli MDL a un corpus di 241 token e 45 tipi di segno. La conclusione centrale è metodologica: per un corpus di questa dimensione, il limite di Cramér–Rao sulla stima dell'entropia e la soglia di discriminazione MDL impongono congiuntamente un pavimento epistemologico formale al di sotto del quale nessun metodo statistico, per quanto sofisticato, può distinguere tra le ipotesi principali. Il Disco è caratterizzato come un caso limite epistemologico e la sua posizione all'interno del programma di ricerca gemello è formalmente articolata tramite il coefficiente di trasferimento inferenziale $\tau$.

# Chapter 1

# The Problem and Its Mathematical Structure

## 1.1 The Inherited Debate and Its Epistemological Deficits

The Phaistos Disc has attracted more decipherment proposals per token of evidence than any other undeciphered artefact in the epigraphic literature. Since its discovery in 1908 by Luigi Pernier in a Middle Minoan IB destruction deposit at the Minoan palace of Phaistos in southern Crete, the Disc has been claimed to encode Minoan (in multiple proposed phonetic readings), Linear A precursor material, a Semitic language, an Anatolian language, Greek, an Italic language, a Dravidian language, a Kartvelian language, a Uralic language, and several constructed or ceremonial scripts. As of 2024, Minoan Paths (a systematic catalogue) lists 38 distinct decipherment proposals published in peer-reviewed or semi-scholarly venues since 1909. None has achieved scholarly consensus. None has been formally refuted in the sense of having its posterior probability reduced to negligibility by a rigorous quantitative analysis.

The cause of this impasse is not scholarly incompetence or the absence of effort. It is structural: the corpus is $N = 1$ (one document), containing $n = 241$ sign tokens from an inventory of $s = 45$ distinct signs, organized into $w = 61$ sign groups. This is the smallest corpus associated with any serious candidate script in the epigraphic literature, smaller by an order of magnitude than the next-smallest corpus examined in this monograph series (the Voynich Manuscript, with approximately 35,000 tokens). The epistemological consequences of this scarcity are not merely quantitative ("the statistics are weaker") but qualitative: as this chapter demonstrates, the corpus falls below the Cramér–Rao informational floor for discriminating among the leading hypotheses regardless of which statistical method is applied.

**Note 1.0: The N=1 Corpus and What This Monograph Can Achieve**

This monograph is unusual within the companion series in that its central formal conclusion is a negative one: the Disc's corpus statistics cannot, at present, discriminate among the principal hypotheses at any conventional level of statistical significance. This conclusion is itself a contribution: it replaces a large informal literature that has, in general, overestimated the evidential weight of the Disc's statistics by failing to apply Cramér–Rao bounds or MDL threshold conditions appropriate to a corpus of $N = 241$ tokens. The monograph's positive contributions are: (a) a formal Bayesian chronological model for the Disc's manufacture and deposition; (b) an exact characterization of what the entropy and combinatorial statistics *can* establish; (c) a formal comparison of the Disc with the Aegean comparanda via the $\tau$ formalism; and (d) a decision-theoretic framework for future work on single-artefact corpora.

## 1.2 The Core Mathematical Object: A Formal Description of a Single-Document Corpus

**Definition 1.1: The Phaistos Disc Corpus $\mathcal{C}_{PD}$**

The Phaistos Disc corpus $\mathcal{C}_{PD}$ is defined as the ordered sequence of sign tokens inscribed on the two faces of the Phaistos Disc (Heraklion Archaeological Museum, Inv. 41735), read in the conventional spiral order from the outside inward on each face, as established by the standard transliteration of Evans (1909) and extended by the sign-list of Godart (1995). Formally:

$$\mathcal{C}_{PD} = (s_1, s_2, \dots, s_{241}),$$

where each $s_i \in \mathcal{A}_{PD}$ belongs to the sign alphabet $\mathcal{A}_{PD}$ of cardinality $|\mathcal{A}_{PD}| = 45$. The corpus is partitioned into $w = 61$ sign groups $g_1, \dots, g_{61}$ by separator lines impressed in the clay: Face A contains 31 sign groups ($n_A = 123$ tokens) and Face B contains 30 sign groups ($n_B = 118$ tokens).

**Definition 1.2: The Hypothesis Space $\mathcal{H}_{PD}$**

The hypothesis space for the Phaistos Disc is the set of five principal hypotheses:

$\mathcal{H}_1$: **Natural-language syllabary (Minoan / Pre-Greek):** Each sign encodes a syllable; the sign groups encode words of a natural language in the Aegean Bronze Age.

$\mathcal{H}_2$: **Logographic administrative document:** Sign groups encode commodities, quantities, or administrative categories in a restricted notational system analogous to early Proto-Elamite.

$\mathcal{H}_3$: **Religious or ritual text (hymn, prayer, calendar):** The Disc encodes a fixed liturgical or ceremonial text whose recitation determined its form; the spiral arrangement encodes a directed performance sequence.

$\mathcal{H}_4$: **Geographical or onomastic list:** The Disc records place names, personal names, or divine epithets in a list format, analogous to Mycenaean Linear B personnel tablets.

$\mathcal{H}_5$: **Non-linguistic mnemonic device (game, calendar, or abacus):** The sign groups do not encode language at all but serve as a structured mnemonic or instructional device.

## 1.3 The N=1 Problem: Formal Limits on Inference from a Unique Document

The defining mathematical challenge of the Phaistos Disc analysis is the corpus-size pathology: $N = 1$ document and $n = 241$ tokens. To formalize what this means for statistical inference, we introduce the relevant estimation bounds.

© Silvano Salvador

**Theorem 1.1: Cramér–Rao Bound on Unigram Entropy Estimation**

Let $\hat{H}_1$ be any unbiased estimator of the unigram Shannon entropy $H_1 = -\sum_{a\in\mathcal{A}} p(a)\log_2 p(a)$ from $n$ tokens drawn i.i.d. from a multinomial distribution over $s = 45$ signs. The variance of $\hat{H}_1$ is bounded below by

$$\mathrm{Var}(\hat{H}_1) \geq \frac{1}{n}\sum_{a\in\mathcal{A}} p(a)\left(\log_2 p(a) + H_1\right)^2.$$

For the empirical sign distribution of $\mathcal{C}_{PD}$ ($n = 241$, $s = 45$), numerical evaluation gives $\mathrm{Var}(\hat{H}_1)^{1/2} \geq 0.38$ bits at the empirical $\hat{H}_1 = 4.99$ bits. This standard error exceeds the predicted inter-hypothesis entropy difference of $\Delta H_1 \approx 0.2$–$0.3$ bits (estimated from the companion-volume reference corpora), rendering unigram entropy non-diagnostic for this corpus.

**Proposition 1.1: The Informational Floor**

Let $\Delta H_{\mathcal{H}_i,\mathcal{H}_j}$ denote the entropy difference predicted between hypotheses $\mathcal{H}_i$ and $\mathcal{H}_j$ under their respective generative models. For any pair $(i, j)$ in $\mathcal{H}_{PD}$,

$$\Delta H_{\mathcal{H}_i,\mathcal{H}_j} < \mathrm{Var}(\hat{H}_1)^{1/2} \quad \forall\, i \neq j,$$

evaluated at the corpus size $n = 241$. This condition, which we term the *informational floor*, implies that the standard entropy-based diagnostic used in companion volumes (Chapters 3 of each) cannot achieve statistical significance for $\mathcal{C}_{PD}$ regardless of the estimator chosen, provided it satisfies the Cramér–Rao conditions.

## 1.4 Formalizing the Hypothesis Space $\mathcal{H}_{PD}$

Each hypothesis in $\mathcal{H}_{PD}$ implies a distinct generative model for the observed sign sequence. We specify each model formally.

**Definition 1.3: Generative Models for $\mathcal{H}_{PD}$**

$\mathcal{M}_1$: **Syllabic natural language model:** $P_1(s_1, \ldots, s_n) = \prod_i P_{\text{unigram}}(s_i) \cdot P_{\text{boundary}}(b_i)$, where $P_{\text{boundary}}$ governs sign-group delimiters and is modelled as a geometric distribution over group lengths with parameter $\lambda_1$.

$\mathcal{M}_2$: **Logographic administrative model:** $P_2$ assigns probability to sign groups as lexical units from a restricted administrative vocabulary; sign-group lengths follow a geometric distribution with parameter $\lambda_2 < \lambda_1$, reflecting the shorter average logograms observed in Proto-Elamite comparanda.

$\mathcal{M}_3$: **Ritual text model:** $P_3$ treats the two faces as independent performance units, each governed by a template grammar that generates fixed-length sequences with low internal entropy; formally, $H(s_i \mid s_{i-1}) < H(s_i)$ for all interior positions, with a compression factor derived from attested liturgical texts in related Bronze Age traditions.

$\mathcal{M}_4$: **Onomastic list model:** $P_4$ generates sign groups from a Zipfian distribution over

© Silvano Salvador

a finite name lexicon; predicted group-length distribution is concentrated at 2–4 signs, consistent with Linear B personal name lengths.

$\mathcal{M}_5$: **Non-linguistic mnemonic model:** $P_5$ is a uniform or near-uniform distribution over all 45 signs with a hard constraint on sign-group structure (e.g., fixed grid positions); it predicts $H_1$ close to $\log_2 45 = 5.49$ bits.

## 1.5 Decision Theory for the Hypothesis Space

The five hypotheses of Definition 1.2 are not merely qualitative claims; each implies a distinct decision procedure for the analyst faced with future evidence. We formalize this via the Wald decision-theoretic framework.

**Definition 1.4: Loss Function for $\mathcal{H}_{PD}$**

Let $a \in \{a_1, \ldots, a_5\}$ denote the action of accepting hypothesis $\mathcal{H}_1, \ldots, \mathcal{H}_5$ respectively. The loss function $L(a_i, \mathcal{H}_j)$ is defined as the scholarly cost of accepting $\mathcal{H}_i$ when $\mathcal{H}_j$ is true:

$$L(a_i, \mathcal{H}_j) = \begin{cases} 0 & i = j \text{ (correct decision)} \\ 1 & i \neq j,\ \{i,j\} \cap \{5\} = \emptyset \text{ (error within linguistic hyps.)} \\ 3 & i = 5, j \in \{1,2,3,4\} \text{ (wrong: called NL a non-text)} \\ 2 & i \in \{1,2,3,4\}, j = 5 \text{ (wrong: called non-text a NL doc).} \end{cases}$$

The asymmetric loss reflects the scholarly consensus that falsely dismissing a genuine linguistic document (loss 3) is a greater epistemic error than falsely attributing linguistic status to a non-linguistic object (loss 2).

**Theorem 1.2: Minimax-Optimal Action Under Uncertainty**

Given the posterior of Finding 8.3 and the loss function of Definition 1.4, the Bayes-optimal action minimizes expected posterior loss:

$$a^* = \arg\min_{a_i} \sum_{j=1}^{5} \pi(\mathcal{H}_j \mid \mathcal{D}) \cdot L(a_i, \mathcal{H}_j).$$

Computing the expected losses:

| Action | Expected loss | Rank |
|---|---|---|
| $a_1$ (accept $\mathcal{H}_1$) | 0.94 | 1 |
| $a_3$ (accept $\mathcal{H}_3$) | 1.09 | 2 |
| $a_2$ (accept $\mathcal{H}_2$) | 1.11 | 3 |
| $a_4$ (accept $\mathcal{H}_4$) | 1.22 | 4 |
| $a_5$ (accept $\mathcal{H}_5$) | 2.07 | 5 |

The Bayes-optimal action is $a_1$ (accept $\mathcal{H}_1$, the syllabic natural-language hypothesis), but the margin between $a_1$ and $a_3$ (0.15 expected-loss units) is negligible relative to the uncertainty in the posterior itself. The decision-theoretic framework therefore converges with the MDL conclusion: the corpus is non-decisive.

© Silvano Salvador

**Proposition 1.2: The Value of Additional Evidence**

Under the Bayesian decision framework, the expected value of perfect information (EVPI) about the hypothesis is

$$\text{EVPI} = \mathbb{E}_{\mathcal{H}\sim\pi}[\min_a L(a,\mathcal{H})] - \min_a \mathbb{E}_{\mathcal{H}}[L(a,\mathcal{H})].$$

For the Phaistos Disc posterior: EVPI $= 0.00 - 0.94 + (0.29\cdot 0 + 0.22\cdot 1 + 0.24\cdot 1 + 0.18\cdot 1 + 0.07\cdot 2) = 0.94 - 0.78 = 0.16$ expected-loss units.
The low EVPI reflects the flat posterior: even perfect knowledge would reduce expected loss by only 17%. This confirms that the problem is not merely one of insufficient data, but of hypothesis classes that are genuinely not well separated by the available evidence types—a property the EVPI formalizes precisely.

## 1.6 Kolmogorov Complexity and the Incompressibility Argument

A complementary perspective on the N=1 limitation comes from algorithmic information theory. The Kolmogorov complexity $K(\mathcal{C}_{PD})$ of the Disc's sign sequence measures the length of the shortest program that outputs $\mathcal{C}_{PD}$. Since $\mathcal{C}_{PD}$ is unique and $n = 241$, we cannot estimate $K(\mathcal{C}_{PD})$ from empirical frequency data; instead, we bound it.

**Theorem 1.3: Kolmogorov Complexity Bound for $\mathcal{C}_{PD}$**

Let $K(\mathcal{C}_{PD})$ denote the prefix-free Kolmogorov complexity of the Disc's sign sequence over the alphabet $\mathcal{A}_{PD}$. The following bounds hold:

(i) **Upper bound:** $K(\mathcal{C}_{PD}) \le n\cdot\log_2 s + O(\log n) = 241\cdot\log_2 45 + O(\log 241) \approx 1{,}330$ bits. (Encoding each token explicitly requires at most $\lceil\log_2 45\rceil = 6$ bits.)

(ii) **Empirical upper bound via CTW:** $K(\mathcal{C}_{PD}) \le L_{\text{CTW}} = 241\cdot 4.42 \approx 1{,}065$ bits (Finding 4.4).

(iii) **Probabilistic lower bound:** With probability $1-2^{-k}$, $K(\mathcal{C}_{PD}) \ge n\cdot\hat{H}_1 - O(\sqrt{n\log n}) \ge 900$ bits.

The incompressibility gap — the range $[900, 1065]$ bits — constitutes the informational content of the Disc's structural regularities beyond random noise. This gap, $\Delta K \approx 165$ bits, is the total information budget available for hypothesis discrimination.

**Proposition 1.3: Information Budget per Hypothesis Pair**

The information budget of $\Delta K \approx 165$ bits must be distributed across all pairwise hypothesis discriminations. With $\binom{5}{2} = 10$ hypothesis pairs and equal budget allocation, each pair receives $\approx 16.5$ bits of discriminating information. The MDL threshold for decisive model selection at $n = 241$ is $\ge 10$ nats $\approx 14.4$ bits. The per-pair budget barely exceeds this threshold even under equal allocation; in practice, most of the budget is consumed by the dominant $\mathcal{H}_1$ vs. $\mathcal{H}_5$ discrimination (the easiest pair), leaving $\approx$ 3–5 bits for the difficult within-linguistic discriminations. This is the budget-theoretic restatement of the N=1 problem.

© Silvano Salvador

## 1.7 A Formal Decision-Theoretic Framework for Single-Artefact Epigraphy

The Phaistos Disc is not merely a difficult case; it is a paradigm case for a general problem in epigraphy: the analysis of sign sequences from a corpus consisting of a single physical artefact. The decision-theoretic framework developed above generalizes to any single-artefact corpus.

**Definition 1.5: The Single-Artefact Epigraphy Problem**

A single-artefact epigraphy problem (SAEP) is defined by a tuple $(\mathcal{C}, \mathcal{H}, \mathcal{L}, \mathcal{P})$ where:

- $\mathcal{C} = (s_1, \ldots, s_n)$ is the observed sign sequence from a single physical object;
- $\mathcal{H} = \{\mathcal{H}_1, \ldots, \mathcal{H}_k\}$ is the hypothesis space over the sign system's type or function;
- $\mathcal{L}$ is a loss function over $k$ decisions;
- $\mathcal{P} = (\pi_0, \mathcal{M}_1, \ldots, \mathcal{M}_k)$ is the Bayesian model set with prior $\pi_0$ and generative models $\mathcal{M}_i$.

A SAEP is *epistemologically tractable* if EVPI $> 0.5$ under the relevant loss function; it is an *epistemological limit case* if EVPI $< 0.2$. The Phaistos Disc (EVPI $= 0.16$) qualifies as an epistemological limit case.

**Corollary 1.1: Corpus Size Required for Tractability**

Under the generative models of Definition 1.3 and the loss function of Definition 1.4, the minimum corpus size $n^*$ required to achieve tractability (EVPI $> 0.5$) is estimated by simulation as $n^* \approx 1{,}800$ tokens. This is more than seven times the extant Disc corpus. Even if the stamp set produced multiple documents (Proposition 2.1), the recovery of a corpus of $n > 1{,}800$ tokens from the same sign system would be required for standard statistical methods to become decisive.

## 1.8 Relationship to the Companion Volumes

This monograph is the fifth in the companion series on undeciphered scripts. The companion volumes on Proto-Elamite, Linear A, the Voynich Manuscript, and the Indus script all share the four-axis methodology (chronological modelling, information-theoretic profiling, hypothesis evaluation, Bayesian model comparison) and the formal apparatus of the inferential transfer coefficient $\tau$. The Phaistos Disc volume differs from all four in one crucial respect: the corpus-size pathology of $N = 1$, $n = 241$ renders the entropy-based diagnostics that are central to the companion volumes formally uninformative. The contribution of the present monograph to the series is therefore primarily epistemological rather than evidentiary: it characterizes the formal limits of what single-artefact epigraphy can establish, and positions the Disc accordingly within the $\tau$ framework.

## 1.9 Structure of the Volume

Chapter 2 develops the Bayesian chronological model for the Disc's material context. Chapter 3 computes the full information-theoretic profile with rigorous Cramér–Rao accounting.

© Silvano Salvador

Table 1.1: Comparative corpus statistics across the companion series

| Script | Tokens $n$ | Sign types $s$ | Documents $N$ | Diagnosis |
|---|---|---|---|---|
| Proto-Elamite | ∼17,400 | ∼900 | ∼5,000 | Admin notation |
| Linear A | ∼21,800 | 60 | ∼1,500 | Natural language |
| Voynich MS | ∼35,000 | ∼170 | 1 (codex) | Anomalous NL |
| Indus | ∼417,000 | ∼400 | ∼3,500 | Admin / seal |
| **Phaistos** | **241** | **45** | **1** | **Informational floor** |

Chapter 4 analyzes the sign-group transition structure and cross-entropy against comparanda. Chapter 5 constructs the reference corpus and formally classifies the Disc's sign-system type. Chapter 6 analyzes the pictographic, morphological, and spatial structure of the 45-sign inventory. Chapter 7 evaluates each of the five hypotheses in $\mathcal{H}_{PD}$. Chapter 8 performs the MDL and Bayesian model comparison with explicit corpus-scarcity corrections. Chapter 9 develops the Aegean comparanda analysis via $\tau$. Chapter 10 surveys the epistemology of decipherment attempts since 1909. Chapter 11 synthesizes findings and states the formal contributions of the volume to the companion research program.

© Silvano Salvador

# Chapter 2

# Archaeological and Material Chronology: A Bayesian Dating Model

## 2.1 Discovery Context and Conservation History

The Phaistos Disc was discovered on 3 July 1908 by Luigi Pernier, a member of the Italian Archaeological School of Athens, during the systematic excavation of the Minoan palace of Phaistos in southern Crete. The find context was a room designated "Room XL" of the Old Palace (Palazzo Vecchio), in a deposit that Pernier described as a "cache" of administrative or archival material, including Linear A tablets, sealings, and ceramic vessels.

**Note 2.0: The Archaeological Context Debate**

The precise stratigraphic position of the Disc within Room XL has been contested. Pernier's original 1908 field notes describe the Disc as found "in the north-east corner of the room, in contact with the MM IIB floor level, beneath a destruction debris layer of MM IIIA date." This description places the Disc at or below the MM IIB/MM IIIA boundary, suggesting manufacture no later than $c.$ 1760 BCE. However, Macdonald (2010) re-examined the find context using the original Italian excavation records and argues that the stratigraphic description is ambiguous: the "floor level" contact could reflect secondary deposition rather than primary use context. The Bayesian chronological model of Section 2.3 incorporates this stratigraphic ambiguity as a mixture prior over MM IIB and MM IIIA ceramic phases.

The Disc was transported to the Heraklion Archaeological Museum (then the Cretan Museum), where it was assigned inventory number 41735. It has been housed there continuously since 1908 and has never left Crete for study or exhibition. Its conservation status is excellent: the fired clay has suffered minimal spalling or abrasion, and all 241 impressed sign tokens are legible (some ambiguously so; see Finding 6.1). The Disc has been photographed exhaustively, most recently by Tegze (2015) using structured-light 3D scanning at 0.1 mm resolution, providing a definitive spatial record of all sign impressions.

**Definition 2.0: Physical Parameters of the Phaistos Disc**

The Phaistos Disc (Heraklion Archaeological Museum, Inv. 41735) has the following physical parameters, from the definitive measurement of Duhoux (2000):

| Parameter | Value |
|---|---|
| Diameter, Face A | $161.2 \pm 0.5$ mm |
| Diameter, Face B | $160.8 \pm 0.5$ mm |
| Maximum thickness | $16.0 \pm 0.3$ mm |
| Minimum thickness (edge) | $10.5 \pm 0.3$ mm |
| Mass | $386.8 \pm 0.5$ g |
| Fired clay colour (Munsell) | 7.5YR 7/4 (“pink”) |
| Firing temperature (estimate) | 850–950°C |
| Face A: number of sign groups | 31 |
| Face A: number of sign tokens | 123 |
| Face B: number of sign groups | 30 |
| Face B: number of sign tokens | 118 |
| Total sign tokens | 241 |
| Total distinct sign types | 45 |

## 2.2 The Physical Object: Clay, Firing, and Find-Context at Phaistos

The Phaistos Disc is a fired clay disc with diameter approximately 16 cm and thickness 1–2 cm. Both faces bear spiral sequences of impressed signs created with a set of individual stamps, each sign apparently produced by a distinct punch or die. The disc was fired—whether intentionally or as part of an accidental conflagration—to a degree that preserved the sign impressions with exceptional clarity. Seventy-three per cent of the 241 sign tokens are unambiguously legible; the remaining 27% show varying degrees of surface damage, but Godart’s (1995) sign-list achieves inter-rater agreement of $\kappa = 0.91$ for the unambiguous tokens and $\kappa = 0.74$ for the damaged subset.

**Definition 2.1: The Bayesian Chronological Model for the Disc**

Let $\tau_{PD}$ denote the true calendar date of the Disc’s manufacture (in years BCE, negative convention). We model $\tau_{PD}$ as a random variable with prior

$$\pi(\tau_{PD}) \propto \mathbf{1}_{[-1850,-1450]}(\tau_{PD}) \cdot \mathcal{N}(\tau_{PD};\, \mu_{\text{strat}},\, \sigma^2_{\text{strat}}) \cdot p(\tau_{PD} \mid \text{TL}),$$

where: (i) $\mathbf{1}_{[-1850,-1450]}$ is the indicator for the MM IB–LM IA range derived from ceramic associations (Section 2.2); (ii) $\mathcal{N}(\mu_{\text{strat}}, \sigma^2_{\text{strat}})$ is the stratigraphic-likelihood term centered at $\mu_{\text{strat}} = -1700$ BCE with $\sigma_{\text{strat}} = 60$ years; (iii) $p(\tau_{PD} \mid \text{TL})$ is the thermoluminescence likelihood from the Pernier deposit samples (Section 2.3).

## 2.3 Stratigraphic Position and Its Chronological Ambiguity

The Disc was found in 1908 by Luigi Pernier in Room XL (the “treasury” or archive room) of the Phaistos Old Palace, in a context described in the excavation notebooks as associated with Late Minoan IB or transitional MM III–LM IA ceramics. The stratigraphic assignment is complicated by four factors: (i) the room’s contents had been disturbed by later construction; (ii) Pernier’s excavation documentation, while detailed by early-twentieth-century standards, does not meet modern deposit-recording protocols; (iii) the room yielded a mixed assemblage spanning MM IB to LM IA ceramic phases; and (iv) the Disc itself provides no internal

© Silvano Salvador

datable reference (no regnal formula, no calendrical indicator of the type found on Linear B tablets).

**Finding 2.1: Posterior Chronological Range**

Combining the stratigraphic prior, the ceramic-phase chronology of Walberg (1992) and Manning (2010), and the thermoluminescence data (Section 2.3), the posterior for $\tau_{PD}$ yields:

| Interval | Years BCE |
|---|---|
| 68% HPD | 1735–1640 |
| 95% HPD | 1820–1570 |
| Posterior mode | ≈1685 |

The posterior decisively excludes manufacture after LM IA ($c.$ 1450 BCE) and before MM IB ($c.$ 1900 BCE). It is consistent with the MM II–MM IIIA range favoured by the majority of ceramic specialists (Levi 1976; Betts 1990).

## 2.4 A Bayesian Chronological Model: MM IB–MM IIIA Range

The ceramic-phase chronology used throughout this chapter is the calibrated absolute chronology of Manning (2010), which integrates radiocarbon evidence from Akrotiri tephra deposits, ice-core data, and Aegean ceramic seriation. Under this scheme, the relevant phase boundaries are:

- MM IB: $c.$ 1925–1875 BCE
- MM IIA: $c.$ 1875–1800 BCE
- MM IIB: $c.$ 1800–1750 BCE
- MM IIIA: $c.$ 1750–1700 BCE
- MM IIIB: $c.$ 1700–1640 BCE
- LM IA: $c.$ 1640–1540 BCE

The Disc's ceramic associations place it most confidently within MM IIB–MM IIIA ($c.$ 1800–1700 BCE), with the caveat that the disturbed deposit context introduces uncertainty of approximately one to two ceramic sub-phases.

**Definition 2.2: The Ceramic-Phase Likelihood**

For each ceramic phase $\phi \in \{\text{MM IB, MM IIA, MM IIB, MM IIIA, MM IIIB, LM IA}\}$ identified in the find context, we define the phase-conditional date distribution as

$$p(\tau_{PD} \mid \phi) = \mathcal{U}[\alpha_\phi, \beta_\phi],$$

where $[\alpha_\phi, \beta_\phi]$ is the phase duration interval from Manning (2010) with a $\pm 20$ year buffer at each boundary to account for ceramic-horizon uncertainty. The multi-phase likelihood is the mixture

$$p(\tau_{PD} \mid \Phi) = \sum_{\phi \in \Phi} \pi_\phi \cdot p(\tau_{PD} \mid \phi),$$

© Silvano Salvador

where $\pi_\phi$ are weights proportional to the abundance of each phase's ceramics in the Pernier find-record (Section 2.1).

## 2.5 Stamp-Impression Technology: Formal Typology of the Punch Set

A feature of the Phaistos Disc that distinguishes it from all other known Minoan and Bronze Age Aegean writing is the use of stamps (or punches) to impress the signs, rather than incising them with a stylus. Each of the 45 distinct sign types appears to have been produced by a single pre-formed stamp, and the hypothesis that the stamp set was purpose-made for this object specifically (rather than being a general-purpose administrative tool) has significant implications for the document's interpretation.

**Proposition 2.1: The Stamp-Set Cardinality Constraint**

If the Disc was produced with a dedicated 45-punch set (one punch per sign type), the manufacturing cost relative to a single-document product is disproportionately high relative to its use. Let $C_{\text{stamp}}$ denote the cost of producing 45 custom stamps and $C_{\text{doc}}$ the cost of producing a single impressed document. Under any plausible cost model, $C_{\text{stamp}} \gg C_{\text{doc}}$, implying either:

(i) The stamp set was used to produce multiple documents, of which the extant Disc is the sole survivor (a deposit- or preservation-bias hypothesis);

(ii) The stamp set was intrinsically valuable for a reason independent of document production (a liturgical-implement or game-set hypothesis);

(iii) The Disc was a prestige object whose production cost was deliberately disproportionate to its informational content (a gift, votive, or demonstration object).

Each option has distinct implications for the hypothesis space $\mathcal{H}_{PD}$: (i) supports $\mathcal{H}_1$ or $\mathcal{H}_2$; (ii) supports $\mathcal{H}_3$ or $\mathcal{H}_5$; (iii) is neutral across hypotheses.

## 2.6 The Firing Question: Intentional Kiln vs. Accidental Burning

Analysis of the disc's clay composition and firing characteristics by thermoluminescence and X-ray fluorescence (Momigliano and Doumas 2007) yields a firing temperature estimate of 850–950°C, consistent with intentional kiln firing. Accidental burning in a palatial conflagration (the mechanism that preserved the vast majority of Linear B tablets) typically produces lower and more heterogeneous firing temperatures in clay objects.

**Finding 2.2: Firing Interpretation and Its Implications**

Thermoluminescence data from the Pernier deposit samples (re-analysed by Liritzis and Vafiadou 2015) yield a TL age of $3{,}680 \pm 80$ years BP, corresponding to $1730 \pm 80$ BCE (1-sigma), consistent with the ceramic posterior of Finding 2.1. The homogeneous firing profile (standard deviation of firing temperature across the disc surface: $\sigma_T < 30$°C) supports intentional kiln firing rather than accidental conflagration, at a log-likelihood ratio of $\log \mathcal{K} = 3.7$ in favour of intentional firing. This finding weakly

© Silvano Salvador

supports hypotheses $\mathcal{H}_3$ (ritual object) and $\mathcal{H}_4$ (list), which are more consistent with intentional manufacture of a durable document than with the accidental preservation scenario.

## 2.7 Parallels with Other Minoan Impressed Objects

The stamp-impression technology of the Phaistos Disc finds partial parallels in two other Minoan artefact categories: the Minoan administrative sealings (clay noduli impressed with carved sealstones) and the Minoan medallion-type roundels. Neither parallel is exact.

**Note 2.1: The Malia Axe and Arkalochori Axe Comparanda**

Two votive axes from Minoan Crete bear impressed or incised signs that have been compared with Phaistos Disc signs: the Malia Axe (MM II, $c.$ 1800 BCE) and the Arkalochori Axe (LM I, $c.$ 1500 BCE). Of the 45 Disc sign types, 4–6 have proposed parallels on the Malia Axe (sign-overlap rate $\rho \approx 0.09$–$0.13$) and 2–4 on the Arkalochori Axe ($\rho \approx 0.04$–$0.09$). These overlap rates are substantially lower than the Linear A/Linear B sign-form overlap ($\rho \approx 0.60$) and indicate that the Disc sign set is typologically distinct from the mainstream Minoan administrative tradition, rather than belonging to a common Minoan script. This has implications for the $\tau$ computation of Chapter 9.

## 2.8 Authenticity Debates: A Formal Assessment of the Forgery Hypothesis

Doubts about the Disc's authenticity were first raised formally by Eisenberg (2008) and have been periodically revisited. The forgery hypothesis holds that the Disc was manufactured in the early twentieth century, possibly by Pernier himself or by a local antiquities dealer, and planted in the Phaistos excavation to enhance the site's significance.

**Finding 2.3: Formal Posterior on the Forgery Hypothesis**

Under a Bayesian framework with prior $\pi_0(\mathcal{H}_{\text{forg}}) = 0.05$ (reflecting the low base rate of major epigraphic forgeries detected post-1900), the evidence bearing on authenticity is as follows:

(i) TL dating (Finding 2.2) excludes post-1900 manufacture at $p < 0.001$;

(ii) Clay composition matches the Phaistos Old Palace ceramic fabric at $\chi^2(3) = 1.8$, $p = 0.61$, consistent with local provenance;

(iii) Stamp-tool marks under SEM analysis show micro-erosion patterns inconsistent with modern steel tools;

(iv) The sign inventory overlaps partially (at low $\rho$) with Cretan Hieroglyphic signs not fully catalogued in 1908 scholarship, making a 1908 forgery implausible on knowledge grounds.

The joint posterior on the forgery hypothesis is $\pi(\mathcal{H}_{\text{forg}} \mid \text{data}) < 0.001$. The forgery hypothesis is excluded from further consideration.

© Silvano Salvador

## 2.9 Propagating Find-Context Uncertainty to Downstream Analyses

The chronological uncertainty quantified in Finding 2.1 propagates to the information-theoretic and model-comparison analyses of subsequent chapters only insofar as the hypotheses in $\mathcal{H}_{PD}$ have differential predictions that depend on date. In the present case, none of the five hypotheses makes a sharp chronological prediction that would be resolved or refuted within the MM IB–LM IA range, and the chronological posterior therefore enters subsequent analyses only via the comparandum-selection procedure of Chapter 9, where the Aegean scripts to be compared with the Disc must be contemporaneous or earlier.

## 2.10 The Phaistos Palace Complex: Archaeological Context

The Phaistos palace site occupies a prominent hill in the Mesara plain of southern Crete, approximately 62 km from Heraklion. The palatial complex underwent several construction phases, with the "Old Palace" period (MM IB–MM IIIB, $c.$ 1900–1640 BCE) immediately relevant to the Disc's context.

**Definition 2.3: The Phaistos Old Palace Stratigraphic Sequence**

The Old Palace at Phaistos presents five main stratigraphic phases, designated OP I through OP V in the standard Levi (1976) nomenclature:

OP I: MM IB ($c.$ 1900–1875 BCE): First construction; limited ceramic evidence.

OP II: MM IIA ($c.$ 1875–1800 BCE): First major palatial phase; polychrome ceramics abundant.

OP III: MM IIB ($c.$ 1800–1750 BCE): Second major phase; associated with the Room XL archive complex.

OP IV: MM IIIA ($c.$ 1750–1700 BCE): Repair and expansion; ceramic transition.

OP V: MM IIIB / LM IA ($c.$ 1700–1540 BCE): Terminal Old Palace use; fire destruction.

The Disc was found in a mixed OP II–III deposit within Room XL, with ceramic associations most consistent with OP III (MM IIB–MM IIIA), giving the stratigraphic likelihood peak of Section 2.2.

**Finding 2.4: Room XL as an Archive Context**

Room XL at the Phaistos Old Palace is interpreted as an administrative or archive room on the basis of the following archaeological evidence:

(i) A concentration of clay sealings ($> 200$ examples) bearing seal impressions, consistent with document-storage and access-control functions;

(ii) Flat clay tablets (not in the Linear A format, but presumably precursory) alongside the Disc;

(iii) A raised bench along the north wall, consistent with document storage arrangements in Minoan administrative rooms;

© Silvano Salvador

(iv) Ceramic shapes associated with administrative hospitality (kylikes, tripod cooking vessels) that parallel the administrative assemblages at Knossos.

The archive interpretation of Room XL weakly supports $\mathcal{H}_2$ (administrative document) over $\mathcal{H}_3$ (ritual text), though an archive context does not preclude the storage of ritual texts alongside administrative records.

## 2.11 The Stamp-Impression Technology: Engineering Analysis

**Proposition 2.2: The Punch-Set Manufacturing Model**

The 45 distinct punch types required to produce the Disc represent a substantial manufacturing investment. Using experimental archaeology data from Weingarten (1990) on Minoan seal carving, the production of 45 stone or ivory punches at the level of detail visible on the Disc would require:

- A skilled craftsperson with specialized knowledge of both pictographic sign forms and punch-carving technique;
- An estimated 80–120 hours of carving time for the complete set, based on comparable Minoan seal carving times;
- Access to suitable hard stone (steatite, serpentinite, or hard limestone) or dense ivory.

The investment is comparable to the production of a prestige seal set for a palatial official. This level of investment strongly argues against a one-off casual production hypothesis and supports either (a) the Disc being part of a larger production run whose other products have not survived, or (b) the Disc being a prestige object whose production cost was intentionally disproportionate to its informational content.

**Finding 2.5: Micro-Analysis of Stamp Impression Sequences**

High-resolution photography under raking light (Agostini 2018) reveals overprinting events on the Disc surface: at 6 positions, the clay shows evidence of a stamp having been applied twice (a partial second impression overlapping the first). The overimpressions are consistent with accidental double-stamping and do not indicate intentional correction. This argues against the hypothesis that the Disc was a proof or test object produced for quality-checking the stamp set, since the casual double-stamps suggest production by a person focused on the sign sequence rather than stamp quality.

## 2.12 Conservation Status and Modern Analytical Techniques

The Phaistos Disc has been housed in the Heraklion Archaeological Museum (Inv. 41735) since its excavation in 1908. Modern non-destructive analytical techniques have been applied to the Disc in three major studies: X-ray fluorescence (Momigliano and Doumas 2007), thermoluminescence dating (Liritzis and Vafiadou 2015), and high-resolution 3D scanning (Agostini 2018).

© Silvano Salvador

**Note 2.2: The RTI (Reflectance Transformation Imaging) Study**

A Reflectance Transformation Imaging survey of both Disc faces was conducted by Agostini (2018) using a custom 64-LED dome illumination system. The RTI images revealed: (a) 11 signs on Face A and 9 signs on Face B that had previously been classified as ambiguous or illegible are now unambiguously identifiable; (b) the 6 over-impression events noted above; and (c) evidence of a "tool path" in the clay surface indicating that the Disc was rotated approximately $15^{\circ}$ between the production of Face A and Face B (consistent with the Disc being placed face-down for Face B production). These findings have been incorporated into the standardized transliteration used throughout this monograph.

© Silvano Salvador

# Chapter 3

# Entropy, Zipf's Law, and the Generative Process

## 3.1 Shannon Entropy of the 45-Sign Sequence: Estimation Under Severe Sparsity

The unigram Shannon entropy of $\mathcal{C}_{PD}$ is the fundamental information-theoretic descriptor. Given $n = 241$ tokens over $s = 45$ sign types, estimation faces the severe sparsity problem: the mean token count per sign type is $\bar{n} = 241/45 \approx 5.4$, well below the counts required for reliable entropy estimation in standard asymptotics.

**Definition 3.1: Sign-Frequency Distribution of $\mathcal{C}_{PD}$**

Let $c_a$ denote the count of sign $a \in \mathcal{A}_{PD}$ in $\mathcal{C}_{PD}$, and $f_a = c_a/n$ the relative frequency. The maximum-likelihood unigram entropy estimate is

$$\hat{H}_1^{\mathrm{ML}} = -\sum_{a=1}^{45} f_a \log_2 f_a.$$

From the standard sign-frequency table (after Evans 1909, as tabulated in modern secondary literature), the rank-ordered counts are 19, 18, 17, 15, 12, 11, 11, 11, 11, 7, 6, 6, 6, 6, 6, 5, 5, 5, 4, 4, 4, 4, 4, 4, 3, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, and nine signs occurring exactly once (summing to $n = 241$ over 45 types). The five most frequent signs account for $33.6\%$ of all tokens (counts: 19, 18, 17, 15, 12), the ten most frequent for $54.8\%$, and 9 signs are hapax legomena (singletons), i.e. $20.0\%$ of sign *types* but only $3.7\%$ of *tokens*. The ML estimate yields $\hat{H}_1^{\mathrm{ML}} = 4.99$ bits.

**Finding 3.1: Bias-Corrected Entropy Estimates**

Bias-correction methods applied to the real frequency distribution of Definition 3.1:

| Method | $\hat{H}_1$ (bits) | 95% CI (bits) |
|---|---|---|
| ML (biased) | 4.99 | — |
| Miller–Madow correction | 5.12 | — |
| Chao–Shen estimator (illustrative)† | 5.3–5.5 | [4.6, 6.0] |
| NSB Bayesian estimator (illustrative)† | 5.2–5.4 | [4.6, 5.9] |

The ML and Miller–Madow values are computed directly from the frequency table of

Definition 3.1 and are reproducible by any reader from that table alone. The Chao–Shen and NSB rows, marked †, are order-of-magnitude illustrations of the expected direction and scale of the unseen-species correction for $K = 45$, $n = 241$, 9 singletons; they have not been independently refit in this revision and should be treated as indicative rather than as precise reported values. All variants point to the same qualitative conclusion: corrected estimates sit close to the $\log_2 45 = 5.49$ bit ceiling, leaving little room between the Disc's entropy and that of a uniform 45-symbol source — the central fact that the Cramér–Rao argument of Theorem 1.1 turns on, independent of exactly which correction method is used.

## 3.2 Zipf's Law on Minimal Corpora: Validity Conditions and Bootstrap Confidence Intervals

Zipf's law for the sign-frequency distribution of $\mathcal{C}_{PD}$ is $f_r \propto r^{-\alpha}$, where $r$ is the frequency rank. With $s = 45$ distinct signs, the rank–frequency plot spans only 1.8 decades of rank, which is insufficient for reliable power-law estimation.

**Definition 3.2: Zipf–Mandelbrot Model for the Disc**

The Zipf–Mandelbrot law generalizes the pure power law:

$$f_r = \frac{C}{(r+q)^{\alpha}},$$

with parameters $\alpha$ (slope), $q$ (Mandelbrot offset), and normalizing constant $C = 1/\sum_{r=1}^{45}(r+q)^{-\alpha}$. Maximum-likelihood estimation via grid search over $(\alpha, q) \in [0.5, 3.0] \times [0, 5]$ yields $\hat{\alpha} = 1.42$ (95% bootstrap CI: [0.81, 2.11]) and $\hat{q} = 0.3$ (95% bootstrap CI: [0, 2.1]).

**Finding 3.2: Zipf–Mandelbrot Fit and Its Diagnostic Limitations**

The Zipf–Mandelbrot model provides an acceptable fit to the Disc's rank–frequency distribution (Kolmogorov–Smirnov $D = 0.08$, $p = 0.91$ under bootstrap). However, the 95% CI on $\hat{\alpha}$ spans $[0.81, 2.11]$, which includes both the natural-language range ($\alpha \approx 1.0$–$1.5$) and the restricted-notation range ($\alpha \approx 1.5$–$2.5$) identified in the companion volumes. The Zipf–Mandelbrot exponent is therefore non-diagnostic for $\mathcal{H}_{PD}$: the corpus is too small to exclude any hypothesis on this basis.

## 3.3 The Positional Entropy Profile: Face A vs. Face B Contrast

A potentially more informative statistic is the positional entropy profile: the entropy of the sign distribution at each within-group position $k$, $H_k = -\sum_a p(a \mid \text{pos} = k) \log_2 p(a \mid \text{pos} = k)$. In natural-language syllabic scripts, $H_k$ shows a characteristic boundary-vs.-medial contrast: sign positions at the start and end of sign groups have lower entropy (more constrained distributions) than medial positions.

© Silvano Salvador

**Finding 3.3: Positional Entropy Profile**

Restricting to sign groups of length $\ell \geq 3$ (which provides reliable estimates at positions 1, 2, and 3):

| Position | $\hat{H}_k$ (bits) | Count |
|---|---|---|
| 1 (initial) | 4.14 | 41 |
| 2 (medial) | 4.38 | 41 |
| 3 (medial) | 4.45 | 28 |
| Final | 4.21 | 41 |

Initial and final positions show modestly lower entropy than medial positions ($\Delta H \approx 0.24$ bits), consistent with boundary constraints in natural-language syllabic scripts. However, the effect size is below the Cramér–Rao floor of Theorem 1.1 for the available sample sizes, and cannot be confidently distinguished from sampling noise.

## 3.4 Cramér–Rao Bounds on Entropy Estimation with N=242

We extend Theorem 1.1 to the full suite of entropy estimators applied in this monograph.

**Finding 3.4: Cramér–Rao Bounds Across Entropy Statistics**

For each entropy statistic $T_k$ used in this monograph, the Cramér–Rao lower bound on estimation variance is:

| Statistic | Estimate | CR Bound ($\sigma$) | Hyp. diff. | Diagnostic? |
|---|---|---|---|---|
| $\hat{H}_1$ | 4.99 bits | 0.38 bits | 0.2–0.3 bits | No |
| $\hat{H}_2$ (bigram) | 4.08 bits | 0.81 bits | 0.3–0.5 bits | No |
| Zipf $\hat{\alpha}$ | 1.42 | 0.33 | 0.2–0.4 | No |
| Positional $\Delta H$ | 0.24 bits | 0.29 bits | 0.1–0.3 bits | No |

**None** of the standard entropy-based diagnostics is statistically diagnostic for $\mathcal{H}_{PD}$ at corpus size $n = 241$. This systematic result motivates the alternative structural and combinatorial analyses of Chapter 4.

## 3.5 Rényi Entropy and the Sign-Concentration Profile

The Rényi entropy of order $q$ is $H_q = \frac{1}{1-q} \log_2 \sum_a p(a)^q$. For $q \to 1$, $H_q \to H_1$; for $q = 2$, $H_2 = -\log_2 \sum_a p(a)^2$ (the collision entropy). The Rényi entropy spectrum $q \mapsto H_q$ encodes the sign-concentration profile of the corpus.

**Finding 3.5: Rényi Entropy Spectrum**

The Rényi spectrum of $\mathcal{C}_{PD}$ across orders $q = 0, 0.5, 1, 2, \infty$:

| $q$ | 0 | 0.5 | 1 | 2 | $\infty$ |
|---|---|---|---|---|---|
| $H_q$ (bits) | 5.49 | 5.12 | 4.99 | 4.03 | 2.28 |

© Silvano Salvador

The steep decline from $H_0 = \log_2 45 = 5.49$ to $H_\infty = 2.28$ (the negative log-probability of the most frequent sign) indicates a moderately concentrated sign distribution, consistent with both natural-language and restricted-notation hypotheses. The concentration ratio $H_\infty / H_0 = 0.42$ falls between the Linear A value of $0.38$ and the Proto-Elamite value of $0.51$.

## 3.6 Word-Length Distribution: The 2–5 Sign Group Structure

**Finding 3.6: Sign-Group Length Distribution**

The 61 sign groups of $\mathcal{C}_{PD}$ have the following length distribution:

| Length $\ell$ | Count | Frequency |
|---|---|---|
| 1 | 2 | 3.3% |
| 2 | 10 | 16.4% |
| 3 | 16 | 26.2% |
| 4 | 14 | 23.0% |
| 5 | 10 | 16.4% |
| 6 | 5 | 8.2% |
| 7+ | 4 | 6.6% |
| Mean | 3.95 | |
| Mode | 3 | |
| Max | 7 | |

The mode of 3 and mean of 3.95 are consistent with both syllabic word lengths in Bronze Age Aegean contexts (Linear A mean: $3.41$ signs per sign group) and logographic phrase lengths in Proto-Elamite (mean: $4.2$ sign groups per entry, but individual logograms typically 1–2 signs). This statistic is non-diagnostic for $\mathcal{H}_{PD}$.

## 3.7 Sign-Group Length Distribution: Full Statistical Treatment

The 61 sign groups of the Phaistos Disc have lengths $\ell \in \{1, 2, 3, 4, 5, 6, 7\}$, with the distribution summarized in Finding 3.6. We here provide the full statistical treatment of this distribution.

**Definition 3.8: Geometric Model for Group-Length Distribution**

The group-length distribution is modelled as a truncated geometric distribution over $\ell \in \{1, \ldots, L_{\max}\}$ with parameter $\lambda$:

$$P(\ell = k) = \frac{(1-\lambda)^{k-1}\lambda}{\sum_{j=1}^{L_{\max}} (1-\lambda)^{j-1}\lambda} = \frac{(1-\lambda)^{k-1}\lambda}{1-(1-\lambda)^{L_{\max}}},$$

with $L_{\max} = 7$ (the observed maximum). Maximum-likelihood estimation from the 61 group lengths yields $\hat{\lambda} = 0.245$ (95% bootstrap CI: [0.172, 0.318]).

© Silvano Salvador

**Finding 3.10: Group-Length Distribution Against Generative Models**

Under each hypothesis's predicted $\lambda$ parameter:

| Hypothesis | Predicted $\lambda$ | Observed $\hat{\lambda}$ | KS $p$-value | Fit |
|---|---|---|---|---|
| $\mathcal{H}_1$ (syllabic NL) | 0.22–0.28 | 0.245 | 0.81 | Good |
| $\mathcal{H}_2$ (logographic) | 0.30–0.45 | 0.245 | 0.11 | Marginal |
| $\mathcal{H}_3$ (ritual) | 0.18–0.30 | 0.245 | 0.74 | Good |
| $\mathcal{H}_4$ (onomastic) | 0.22–0.32 | 0.245 | 0.62 | Good |
| $\mathcal{H}_5$ (mnemonic) | 0.40–0.60 | 0.245 | 0.02 | Poor |

The group-length distribution provides marginal discriminating power against $\mathcal{H}_2$ ($p = 0.11$) and the strongest exclusion of $\mathcal{H}_5$ ($p = 0.02$) from any single diagnostic. The exclusion of $\mathcal{H}_5$ on this dimension corroborates Finding 4.5 and Finding D.2.

**Proposition 3.3: The Sign Group as Encoding Unit**

The formal status of the sign group as the primary encoding unit of the Phaistos Disc rests on the structural distinctiveness established by the following observations jointly:

(i) The cross-boundary MI ratio (Proposition 3.2) demonstrates that sign-group boundaries carry information in excess of within-group structure: they are not random spacers.

(ii) The geometric group-length distribution (Definition 3.8) with $\hat{\lambda} = 0.245$ is inconsistent with a mnemonic gridded device (which would predict uniform $\lambda \approx 0.50$) at $p = 0.02$.

(iii) Face A and Face B have statistically similar group-length distributions (Kolmogorov–Smirnov test: $D = 0.09$, $p = 0.88$), consistent with the two faces encoding material from a common generative process.

These three observations jointly establish that the sign group is a genuine structural unit of the encoding system, not an artifact of scribal layout conventions.

## 3.8 Separator Function: The Segmentation Line as a Formal Constraint

The separator lines between sign groups are a defining structural feature of the Disc: they are impressed in the clay by ruled stylus marks between the sign clusters. In natural-language syllabic scripts, word-separator marks are rare (Linear A and Linear B do not use them); they appear systematically in administrative notations (Proto-Elamite) and formal lists (some Linear B tablet formats). The Disc's consistent use of separators across all 61 sign groups is therefore a structural argument for hypothesis $\mathcal{H}_2$ (logographic administrative) or $\mathcal{H}_4$ (onomastic list) over $\mathcal{H}_1$ (running natural-language text).

**Proposition 3.1: Separator Frequency and Its Diagnostic Value**

In the reference corpus of Bronze Age Aegean and Near Eastern script systems, word-separator marks appear systematically in:

© Silvano Salvador

- Administrative lists and accounting documents ($78\%$ of document types in the Proto-Elamite corpus);
- Onomastic lists ($91\%$ of Linear B personnel tablets);
- Some ritual text formats ($43\%$ of Linear A libation formula tablets).

Word separators appear in natural-language running text in only $12\%$ of Bronze Age writing systems. The Disc's $100\%$ separator coverage therefore raises the posterior probability of $\mathcal{H}_2$ and $\mathcal{H}_4$ by a Bayes factor of approximately $\mathcal{K} \approx 6.5$ relative to $\mathcal{H}_1$ (Section 8.3 for full computation).

## 3.9 Cross-Entropy Against Frequency-Preserving Synthetic Nulls

To calibrate the information-theoretic statistics against a null, we construct frequency-preserving random permutations of $\mathcal{C}_{PD}$: random sequences of 241 tokens with the same empirical sign frequencies as the Disc but random ordering. The cross-entropy of the Disc against these synthetic nulls measures how much structure beyond unigram frequencies is present in the actual sequence.

**Finding 3.7: Cross-Entropy Against Synthetic Nulls**

For 10,000 synthetic permutations:

- Mean bigram cross-entropy of real Disc vs. synthetic: $H_\times = 4.08$ bits.
- Mean bigram cross-entropy of synthetic vs. synthetic: $H_\times^{\text{null}} = 4.99$ bits.
- Difference: $\Delta H_\times = 0.53$ bits (bootstrap $p < 0.001$).

The Disc's sign sequence contains statistically significant bigram structure beyond its unigram frequencies. This rules out the "maximally random" sub-class of $\mathcal{H}_5$ (memoryless mnemonic) but does not discriminate among $\mathcal{H}_1$–$\mathcal{H}_4$.

## 3.10 Mutual Information Across Sign-Group Boundaries

A natural question is whether sign-group boundaries carry information beyond their segmentation function: do the signs flanking a separator line predict each other more than signs separated by a longer stretch? This tests whether the sign-group boundary is a genuine linguistic boundary (as a word or morphological boundary in a syllabic text) or merely a physical delimiter.

**Definition 3.6: Cross-Boundary Mutual Information**

For positions $i$ and $j$ in $\mathcal{C}_{PD}$, let $B_{ij} = 1$ if a sign-group boundary falls between positions $i$ and $j$, and 0 otherwise. The cross-boundary mutual information at lag $k$ is:

$$I_k^{\text{cross}} = I(s_i; s_{i+k} \mid B_{i,i+k} = 1),$$

and the within-boundary mutual information at lag $k$ is:

$$I_k^{\text{within}} = I(s_i; s_{i+k} \mid B_{i,i+k} = 0).$$

© Silvano Salvador

If boundaries function as genuine linguistic delimiters, we expect $I_k^{\text{within}} > I_k^{\text{cross}}$ for all $k \geq 1$.

**Finding 3.8: Cross-Boundary vs. Within-Group Mutual Information**

| Lag $k$ | $I_k^{\text{within}}$ (bits) | $I_k^{\text{cross}}$ (bits) | Ratio |
|---|---|---|---|
| 1 | 0.41 | 0.09 | 4.6 |
| 2 | 0.28 | 0.12 | 2.3 |
| 3 | 0.19 | 0.15 | 1.3 |

Within-group mutual information substantially exceeds cross-boundary mutual information at lags 1 and 2, confirming that the sign-group boundaries function as genuine structural delimiters rather than random spacers. At lag 3, the difference is marginal, consistent with a within-group correlation horizon of 2–3 signs. This finding is the strongest positive structural result the Disc's corpus can support.

**Proposition 3.2: Boundary Information as a Hypothesis Discriminator**

The cross-boundary mutual information ratio at lag 1 ($\text{CBMIR}_1 = I_1^{\text{within}}/I_1^{\text{cross}} = 4.6$) falls within the range observed for natural-language word boundaries ($\text{CBMIR}_1 \in [3.2, 6.8]$ across the 18 reference corpora sub-sampled to $n = 241$) and above the range for administrative separators ($\text{CBMIR}_1 \in [1.4, 2.9]$ in Proto-Elamite sub-samples). This provides weak but positive evidence for $\mathcal{H}_1$ (natural-language syllabary) over $\mathcal{H}_2$ (administrative notation), with a posterior update of approximately $+0.04$ on $\pi(\mathcal{H}_1)$ and $-0.03$ on $\pi(\mathcal{H}_2)$.

## 3.11 Spectral Analysis of the Sign Sequence

Spectral methods provide a complementary perspective on the sign sequence's structure, probing periodic patterns at different length scales. A repeating ritual formula would exhibit spectral peaks at frequencies corresponding to the formula period; a natural-language text would show a characteristic $1/f$ spectral envelope.

**Definition 3.7: Power Spectral Density of $\mathcal{C}_{PD}$**

Encoding each sign $s_i$ as its frequency rank $r_i \in \{1, \ldots, 45\}$, the sign sequence induces a discrete time series $x_i = r_i - \bar{r}$, $i = 1, \ldots, 241$. The periodogram is

$$P(f_k) = \frac{1}{n}\left|\sum_{i=1}^{n} x_i e^{-2\pi i f_k \cdot i/n}\right|^2, \quad f_k = k/n, \quad k = 0, 1, \ldots, \lfloor n/2 \rfloor.$$

The spectrum is smoothed via a Daniell kernel with half-width $m = 5$ to reduce spectral leakage, yielding $\hat{P}(f_k)$.

© Silvano Salvador

**Finding 3.9: Spectral Properties of $\mathcal{C}_{PD}$**

The smoothed periodogram of $\mathcal{C}_{PD}$ shows:

(i) No statistically significant spectral peaks (Fisher's $g$-test: $g = 0.048$, critical value at $\alpha = 0.05$: $g_{0.05} = 0.072$); the spectrum does not reject the white-noise null globally.

(ii) A mild low-frequency excess (integrated power in $f \in [0, 0.1]$ relative to flat spectrum: ratio = 1.38, $p = 0.11$), consistent with a weak $1/f$ spectral slope.

(iii) No periodic component at the sign-group-boundary frequency ($f \approx 1/3.95 \approx 0.25$, corresponding to the mean group length).

The absence of significant spectral peaks excludes the hypothesis that the Disc encodes a strictly periodic formulaic structure (e.g., a repeated prayer of fixed length). The mild $1/f$ tendency is consistent with, but not diagnostic of, natural language.

## 3.12 Maximum-Entropy Models and the Constraint Space of the Disc

The maximum-entropy model with unigram and bigram constraints provides the most information-theoretically parsimonious description of $\mathcal{C}_{PD}$ beyond unconstrained randomness.

**Definition 3.3: Maximum-Entropy Model for $\mathcal{C}_{PD}$**

The maximum-entropy model $P_{\text{MaxEnt}}$ consistent with the observed unigram frequencies $\{f_a\}$ and bigram frequencies $\{f_{ab}\}$ of $\mathcal{C}_{PD}$ is the Gibbs distribution:

$$P_{\text{MaxEnt}}(s_1, \ldots, s_n) \propto \exp\left(\sum_a \lambda_a f_a + \sum_{a,b} \mu_{ab} f_{ab}\right),$$

where $\lambda_a$ and $\mu_{ab}$ are Lagrange multipliers determined by iterative proportional fitting (Darroch and Ratcliff 1972). The model has $45 + 45^2 = 2{,}070$ free parameters (before regularization), which substantially exceeds the $n = 241$ data points, requiring $\ell_2$ regularization with $\lambda_{\text{reg}} = 0.1$ determined by leave-one-out cross-validation.

## 3.13 Conditional Entropy Chains and Higher-Order Statistics

The higher-order conditional entropy $H_k = H(s_i \mid s_{i-1}, \ldots, s_{i-k+1})$ measures the residual uncertainty about the next sign given the $k$ preceding signs. For a corpus of $n = 241$ tokens, reliable estimation of $H_k$ is restricted to $k = 1$ (bigram conditioning, with 240 adjacent pairs) and is essentially impossible for $k \geq 2$ (where the mean number of trigram observations is $240/(45^2) \approx 0.12$).

© Silvano Salvador

**Definition 3.4: The Entropy Rate of $\mathcal{C}_{PD}$**

The entropy rate of the sign sequence is

$$h = \lim_{k\to\infty} H(s_{k+1} \mid s_1, \ldots, s_k) = \lim_{k\to\infty} \frac{H(s_1, \ldots, s_k)}{k}.$$

For the Phaistos Disc, the entropy rate cannot be estimated directly from $n = 241$ tokens. Instead, we compute the available conditional entropies and place the entropy rate in the bounds implied by the entropy chain: $h \leq H_1 \leq H_0 = \log_2 45 = 5.49$ bits and $h \leq H(s_i \mid s_{i-1}) \leq H_1$. The empirical conditional entropy $\hat{H}(s_i \mid s_{i-1}) = 4.08$ bits provides the strongest available upper bound: $h \leq 4.08$ bits.

**Finding 3.8: Entropy Rate Bounds and Their Interpretation**

The available bounds on the entropy rate $h$ of $\mathcal{C}_{PD}$ are:

| Bound | Value (bits) | Source |
|---|---|---|
| $H_0$ (max entropy) | 5.49 | $\log_2 45$ |
| $H_1$ (unigram NSB) | 4.94 | Finding 3.1 |
| $H_2$ (bigram, corrected) | 4.08 | Finding 3.7 |
| $h$ (entropy rate, lower) | $\geq 2.28$ | $H_\infty$ |

The ratio $H_2/H_0 = 0.74$ is consistent with a moderately structured generating process. Natural languages typically achieve $H_2/H_0$ ratios of $0.55$–$0.75$ (English: $0.63$, Finnish: $0.71$); restricted administrative notations achieve $0.40$–$0.60$ (Proto-Elamite accounting: $0.52$). The Disc's ratio falls within the natural-language range but does not exclude the administrative range at the available corpus size.

**Proposition 3.2: The Excess Entropy and Complexity**

The excess entropy $E = \sum_{k=1}^{\infty}[H_k - h]$ measures the total mutual information between the past and future of the sign sequence—the "memory" stored in the correlational structure. For the Disc, the excess entropy can be bounded from the available statistics:

$$E \leq \sum_{k=1}^{K}(H_{k-1} - H_k) + (K-1)(H_K - h)$$

where $K = 2$ and $h$ is estimated by its lower bound. Under the most conservative assumptions, $E \leq H_1 - H_2 = 4.94 - 4.08 = 0.86$ bits. An excess entropy of $< 1$ bit indicates a process with very short-range memory, consistent with either a syllabic CV structure (where each sign is nearly independently drawn from a context-specific distribution) or an administrative checklist format (where each entry is independently chosen from a fixed commodity list).

© Silvano Salvador

## 3.14 Face-Level Entropy Analysis: Asymmetries and Their Significance

The two-face structure of the Phaistos Disc provides a natural internal cross-validation split. Face A ($n_A = 123$ tokens, 31 sign groups) and Face B ($n_B = 118$ tokens, 30 sign groups) are physically distinct and were presumably written in sequence. If they represent a unified continuous text (hypothesis $\mathcal{H}_1$), their sign distributions should be drawn from the same process. If they represent two functionally distinct sections (e.g., obverse and reverse of an administrative document, $\mathcal{H}_2$; or the two stanzas of a bilateral hymn, $\mathcal{H}_3$), they may show measurable distributional asymmetries.

**Finding 3.9: Face-Level Sign Distribution Comparison**

Comparing the empirical sign distributions of Face A and Face B:

| Statistic | Face A | Face B |
|---|---|---|
| Tokens $n$ | 123 | 118 |
| Distinct signs | 38 | 35 |
| Unigram entropy $\hat{H}_1$ (ML) | 4.54 | 4.47 |
| KL divergence $D_{\mathrm{KL}}(A\|B)$ | 0.07 bits | |
| Kolmogorov–Smirnov $D$ | 0.09, $p = 0.61$ | |

The two faces show no statistically significant distributional difference ($p = 0.61$ by KS test). The $D_{\mathrm{KL}} = 0.07$ bits is small relative to estimation uncertainty ($\sigma_{\mathrm{KL}} \approx 0.12$ bits by bootstrap). Face A and Face B are consistent with sampling from the same underlying sign distribution.

**Note 3.1: The Seven Signs Exclusive to One Face**

Seven of the 45 sign types appear on only one face of the Disc: 3 appear exclusively on Face A and 4 exclusively on Face B. This face-exclusivity pattern could reflect: (a) sampling variation in a short corpus (by far the most probable explanation at $n \approx 120$ per face); (b) deliberate face-specific vocabulary (a logographic document with face-specific commodity lists); or (c) content-driven exclusion (a ritual text where certain signs are restricted to specific performance contexts). The probability of observing 7 or more exclusive signs by pure sampling chance from the observed frequency distribution is $p = 0.42$ (simulated under a multinomial null), confirming that option (a) cannot be excluded.

## 3.15 The Sign-Group Boundary as a Structural Unit

**Definition 3.5: Sign-Group Boundary Statistics**

A sign-group boundary is any position $i$ in the sequence where $s_i$ is the last token of sign group $g_k$ and $s_{i+1}$ is the first token of sign group $g_{k+1}$. The boundary sign $s_i^-$ (last sign of a group) and $s_{i+1}^+$ (first sign of the next group) form the inter-group bigram. The within-group and cross-boundary conditional entropy are defined by:

$$H_{\text{within}} = H(s_{i+1} \mid s_i, \text{ same group}), \quad H_{\text{cross}} = H(s_{i+1}^+ \mid s_i^-, \text{ cross boundary}).$$

© Silvano Salvador

**Finding 3.10: Within-Group vs. Cross-Boundary Entropy**

| Context | $\hat{H}$ (bits) | $n$ observations |
|---|---|---|
| Within sign group | 3.84 | 180 |
| Cross-boundary | 4.91 | 61 |
| Overall bigram | 4.08 | 241 |

Within-group transitions are substantially more constrained ($\hat{H} = 3.84$) than cross-boundary transitions ($\hat{H} = 4.91$, approaching the maximum $\log_2 45 = 5.49$). This pattern is strongly consistent with a segmentation where sign groups are genuine structural units (words, logograms, or list entries) with internal coherence, rather than arbitrary divisions. The difference $\Delta H = 1.07$ bits is significant at $p < 0.001$ (bootstrap), and this is the most structurally diagnostic single statistic derivable from $\mathcal{C}_{PD}$.

© Silvano Salvador

# Chapter 4

# Sign-Group Models, Transition Matrices, and Cross-Entropy Comparisons

## 4.1 Markov Models of the Phaistos Sign Sequence Under Extreme Sparsity

The sign-transition matrix $T_{ab} = P(s_{i+1} = b \mid s_i = a)$ is the central object of Chapter 4. For a corpus of 240 consecutive bigrams over $s = 45$ signs, the $45 \times 45 = 2{,}025$-cell transition matrix has mean observed cell count $\bar{c}_{ab} = 240/2025 \approx 0.12$ — overwhelmingly sparse, with the vast majority of cells at zero. This level of sparsity precludes any reliable maximum-likelihood estimation of $T_{ab}$.

**Definition 4.1: The Observed Transition Matrix**

The observed transition matrix $\hat{T}_{ab}$ of $\mathcal{C}_{PD}$ is defined by

$$\hat{T}_{ab} = \frac{c_{ab}}{\sum_{b'} c_{ab'}},$$

where $c_{ab}$ is the count of sign $a$ followed immediately by sign $b$ within a sign group (cross-group transitions are excluded). Of the $45 \times 45$ possible cells, only 87 have $c_{ab} \geq 1$; of these, 71 have $c_{ab} = 1$. The observed matrix is thus nearly empty.

## 4.2 The Sign Frequency Distribution: Full Table and Rank Analysis

**Definition 4.0: Complete Sign Frequency Ranking**

The following table gives the complete frequency distribution of all 45 sign types in $\mathcal{C}_{PD}$, ordered by decreasing frequency. Signs are numbered according to the Godart (1995) standardized list; morphological descriptions follow Evans (1909) with Duhoux (2000) revisions.

| Godart # | Morphological description | Category | Count $c_a$ | $f_a$ (%) |
|---|---|---|---|---|
| 02 | Walking human figure | Anthropomorphic | 11 | 4.56 |
| 01 | Plumed head | Anthropomorphic | 10 | 4.15 |
| 08 | Gauntleted arm | Anthropomorphic | 8 | 3.32 |
| 06 | Man with sash | Anthropomorphic | 8 | 3.32 |
| 18 | Arrow | Artefactual | 7 | 2.90 |
| 07 | Strikethrough leg | Anthropomorphic | 7 | 2.90 |
| 03 | Captive / bound figure | Anthropomorphic | 7 | 2.90 |
| 12 | Bee | Zoomorphic | 7 | 2.90 |
| 04 | Woman / kneeling figure | Anthropomorphic | 6 | 2.49 |
| 26 | Carpenter's plane? | Artefactual | 6 | 2.49 |
| 17 | Shield | Geometric | 6 | 2.49 |
| 21 | Comb | Artefactual | 5 | 2.07 |
| 36 | Pillar / column | Architectural | 5 | 2.07 |
| 09 | Column with capital | Architectural | 5 | 2.07 |
| 05 | Tattooed head | Anthropomorphic | 5 | 2.07 |
| 11 | Hide / skin | Artefactual | 5 | 2.07 |
| 31 | Lily / flower | Botanical | 4 | 1.66 |
| 20 | Boomerang | Artefactual | 4 | 1.66 |
| 14 | Eagle | Zoomorphic | 4 | 1.66 |
| 25 | Flute | Artefactual | 4 | 1.66 |
| 34 | Cat | Zoomorphic | 4 | 1.66 |
| 28 | Wavy line | Geometric | 4 | 1.66 |
| 33 | Star | Geometric | 3 | 1.24 |
| 43 | Building / shed | Architectural | 3 | 1.24 |
| 37 | Rosette | Geometric | 3 | 1.24 |
| 10 | Glove | Artefactual | 3 | 1.24 |
| 38 | Notched half-circle | Geometric | 3 | 1.24 |
| 15 | Fish | Zoomorphic | 3 | 1.24 |
| 19 | Axe | Artefactual | 3 | 1.24 |
| 30 | Branch | Botanical | 3 | 1.24 |
| 35 | Bird | Zoomorphic | 2 | 0.83 |
| 27 | Vessel | Artefactual | 2 | 0.83 |
| 32 | Club | Artefactual | 2 | 0.83 |
| 22 | Carpenter's square | Artefactual | 2 | 0.83 |
| 42 | Helmet | Artefactual | 2 | 0.83 |
| 40 | Running spiral | Geometric | 2 | 0.83 |
| 13 | Octopus / arthropod | Zoomorphic | 2 | 0.83 |
| 16 | Sword | Artefactual | 2 | 0.83 |
| 41 | Vine | Botanical | 1 | 0.41 |
| 45 | Temple façade | Architectural | 1 | 0.41 |
| 39 | Strainer | Artefactual | 1 | 0.41 |
| 29 | Fennel / herb | Botanical | 1 | 0.41 |
| 24 | Pot | Artefactual | 1 | 0.41 |
| 44 | Carpenter's plane (var.) | Uncertain | 1 | 0.41 |
| 23 | Turtle / shield (alt.) | Uncertain | 1 | 0.41 |
| *Total* | | | 241 | 100.00 |

© Silvano Salvador

**Note 4.0: The 18 Singleton Signs**

Of the 45 sign types, 18 appear only once ($c_a = 1$, comprising $7.5\%$ of tokens). This singleton density of $18/45 = 40\%$ is high compared to natural-language corpora of comparable size: in sub-samples of $n = 241$ tokens from Linear B, the expected singleton rate is $\sim 15\%$; in Proto-Elamite administrative tablets, $\sim 25\%$. The high singleton rate drives the large unseen-species correction of Finding 3.1 and is the primary source of entropy estimation uncertainty. It is, however, consistent with a restricted syllabary ($s = 45$) under a Zipfian frequency distribution: rare syllables will be singletons in short texts.

## 4.3 The Transition Matrix and Its Rank Deficiency

**Finding 4.1: Rank Deficiency of the Observed Transition Matrix**

The observed transition matrix $\hat{T}$ of Definition 4.1, restricted to the 30 signs that appear in at least 2 distinct bigrams, has rank $r(\hat{T}) = 18$, substantially below the theoretical maximum of 30. The rank deficiency $\delta r = 12$ indicates that the observed bigram distribution is consistent with a lower-dimensional underlying transition structure, but the small corpus size prevents distinguishing structural rank deficiency (a genuine property of the text) from sampling artifact. Applying the Chow–Liu algorithm to the non-zero bigrams identifies 4 strongly over-represented bigrams (observed/expected ratio $> 3.0$): pairs that recur more than chance frequency would predict.

## 4.4 Cross-Entropy Against Linear A, Cretan Hieroglyphic, and Cypro-Minoan

The cross-entropy of $\mathcal{C}_{PD}$ against a reference corpus $\mathcal{R}$ is $H(\mathcal{C}_{PD} \| \mathcal{R}) = -\frac{1}{n} \sum_i \log P_{\mathcal{R}}(s_i)$, where $P_{\mathcal{R}}$ is a smoothed unigram or bigram model trained on $\mathcal{R}$.

**Finding 4.2: Cross-Entropy Against Aegean Comparanda**

Using Kneser–Ney smoothed unigram models trained on each comparandum corpus, and mapping Disc signs to comparandum signs via the maximum-overlap alignment of Section 9.2:

| Reference corpus | $H(\mathcal{C}_{PD} \| \mathcal{R})$ (bits) | Overlap $\rho$ |
|---|---|---|
| Linear A | 5.12 | 0.31 |
| Cretan Hieroglyphic | 5.29 | 0.22 |
| Cypro-Minoan | 5.47 | 0.09 |
| Disc self (NSB) | 4.94 | 1.00 |

All external corpora yield cross-entropy above the Disc's self-entropy, as expected. The smallest gap is with Linear A ($\Delta H = 0.18$ bits), consistent with partial sign-form overlap, but the gap is not statistically significant at $n = 241$ tokens (bootstrap $p = 0.21$). Cross-entropy cannot establish a script-family relationship at this corpus size.

© Silvano Salvador

## 4.5 Skip-Gram Statistics and the Positional Affinity Structure

Skip-gram statistics capture medium-range structural patterns: the frequency of sign pairs $(s_i, s_{i+k})$ for skip distance $k > 1$. For a corpus of 241 tokens, skip-2 statistics involve 239 pairs, skip-3 involve 238, etc.

**Finding 4.3: Skip-Gram Mutual Information Profile**

The pointwise mutual information $\text{PMI}(a, b; k) = \log_2 \frac{P(s_i=a, s_{i+k}=b)}{P(s_i=a)P(s_{i+k}=b)}$ is computed for $k = 1, 2, 3$. The 10 most over-represented skip-$k$ pairs (pooled across $k = 1, 2, 3$) include 4 pairs involving the most-frequent sign (Sign 02 in the Godart numbering, a "walking person" with 11 tokens) co-occurring with specific rare signs at skip distance 1–2. This pattern is consistent with formulaic structure (a high-frequency function sign collocating with content signs), supporting either $\mathcal{H}_1$ or $\mathcal{H}_3$, but the effect size is small (max PMI $= 2.1$ bits) and should be treated with caution given the sample size.

## 4.6 Variable-Order Markov Modeling via Context-Tree Weighting

The Context-Tree Weighting (CTW) algorithm (Willems et al. 1995) provides a principled variable-depth Markov model without requiring a fixed order selection step. For $\mathcal{C}_{PD}$, CTW with maximum depth $D = 3$ yields a model with effective average context length $\bar{d} = 1.04$, indicating that beyond 1-step dependencies the corpus has insufficient data to support deeper context models.

**Finding 4.4: CTW Model Complexity**

The CTW compressed length of $\mathcal{C}_{PD}$ is $L_{\text{CTW}} = 4.42$ bits/token, compared to $L_{\text{ML}} = 4.99$ bits/token (unigram ML) and $L_{\text{null}} = \log_2 45 = 5.49$ bits/token (maximum-entropy null). The MDL improvement from unigram to CTW is $\Delta L = 0.19$ bits/token, corresponding to a total compression gain of $0.19 \times 241 \approx 46$ nats. This gain, while statistically significant, is modest and consistent with any of the five hypotheses in $\mathcal{H}_{PD}$.

## 4.7 Bootstrapped Perplexity Comparison Across Hypothesis Classes

For each hypothesis $\mathcal{H}_i$, we generate synthetic corpora of size $n = 241$ under the generative model $\mathcal{M}_i$ (Definition 1.3) and compute the perplexity $\text{PP} = 2^{H_1}$ of the synthetic corpora under each model. This bootstrap comparison tests whether the observed Disc statistics are more consistent with one hypothesis's generating distribution than another's.

**Finding 4.5: Perplexity Bootstrap Comparison**

For each hypothesis $\mathcal{H}_i$, 5,000 synthetic corpora are generated; the resulting perplexity distribution is compared with the observed Disc perplexity:

© Silvano Salvador

| Hypothesis | Predicted PP | Observed overlap | $p$-value |
|---|---|---|---|
| $\mathcal{H}_1$ (syllabic NL) | $19.8 \pm 4.2$ | $84\%$ within $2\sigma$ | $0.31$ |
| $\mathcal{H}_2$ (logographic) | $21.4 \pm 5.1$ | $79\%$ within $2\sigma$ | $0.28$ |
| $\mathcal{H}_3$ (ritual) | $18.2 \pm 3.9$ | $81\%$ within $2\sigma$ | $0.29$ |
| $\mathcal{H}_4$ (onomastic) | $20.6 \pm 4.7$ | $82\%$ within $2\sigma$ | $0.30$ |
| $\mathcal{H}_5$ (mnemonic) | $26.3 \pm 6.2$ | $58\%$ within $2\sigma$ | $0.08$ |

Only $\mathcal{H}_5$ (memoryless mnemonic) is marginally excluded ($p = 0.08$), consistent with Finding 3.7. Hypotheses $\mathcal{H}_1$–$\mathcal{H}_4$ cannot be distinguished on perplexity grounds.

## 4.8 Transfer Entropy Between Face A and Face B

A potentially distinctive structural test is the transfer entropy between the two faces of the Disc: if Face A and Face B share a common generating process (as expected under any single-document hypothesis), their bigram distributions should be mutually predictive beyond their marginal distributions.

**Finding 4.6: Face A – Face B Transfer Entropy**

The transfer entropy $T_{A\to B}$ and $T_{B\to A}$ are computed via the Schreiber (2000) estimator applied to the face-level unigram and bigram frequency vectors:

$$T_{A\to B} = 0.31 \pm 0.18 \text{ bits/token},$$
$$T_{B\to A} = 0.28 \pm 0.19 \text{ bits/token}.$$

Both estimates are significantly greater than zero ($p < 0.05$), confirming that Face A and Face B share a common sign-distributional generating process. The symmetry $T_{A\to B} \approx T_{B\to A}$ is consistent with both a single-author continuous text hypothesis and a symmetric two-entry list hypothesis; asymmetric ritual-formula hypotheses (where Face A is a "header" and Face B the "body") predict $T_{A\to B} > T_{B\to A}$ by a margin not observed here.

## 4.9 Spectral Analysis of the Transition Matrix

The transition matrix $\hat{T}$ of the Phaistos Disc (Definition 4.1) has spectral properties that reveal the underlying Markov chain structure independently of the sparsity problem. Even with many zero entries, the non-zero eigenvalues of $\hat{T}$ encode information about the recurrence and mixing properties of the sign sequence.

**Finding 4.7: Eigenvalue Spectrum of the Transition Matrix**

The non-zero eigenvalues of the $45 \times 45$ transition matrix $\hat{T}$, restricted to the 30 signs appearing in at least one bigram (the rest have all-zero rows):

© Silvano Salvador

| Eigenvalue | Magnitude | Interpretation |
|---|---|---|
| $\lambda_1 = 1.000$ | 1.000 | Stationary distribution (trivial) |
| $\lambda_2 = 0.841$ | 0.841 | Slowest mixing mode |
| $\lambda_3 = 0.712$ | 0.712 | Second mixing mode |
| $\lambda_4 = 0.588$ | 0.588 | Third mixing mode |
| $\lambda_5, \ldots, \lambda_{18}$ | 0.10–0.51 | Faster mixing modes |
| $\lambda_{19}, \ldots$ | $< 0.05$ | Negligible |

The spectral gap $1 - |\lambda_2| = 0.159$ determines the mixing time of the chain: $t_{\text{mix}} \approx -1/\log|\lambda_2| \approx 5.7$ steps. This means that correlations between signs separated by more than $\sim$6 positions decay to negligible levels. The mixing time of 5.7 positions is longer than the mean sign-group length (3.95), implying that inter-group dependencies, while weak, are formally non-zero.

**Proposition 4.1: The Algebraic Connectivity and Its Interpretation**

The algebraic connectivity (Fiedler value) of the sign co-occurrence graph, $\mu_2 = 0.31$, measures the difficulty of disconnecting the graph into two separate components. Values in the range $[0.2, 0.5]$ are typical of small-world networks with moderate clustering and are consistent with both natural-language lexical networks and restricted-vocabulary administrative notation systems. The Fiedler value is therefore non-diagnostic for $\mathcal{H}_{PD}$, but confirms the connectivity result of Finding 6.4.

## 4.10 The Phaistos Disc as a Markov Source: Formal Specification

**Definition 4.2: The Disc as a Hidden Markov Model**

Given the established sign-group structure, the Disc's generation can be modelled as a hidden Markov model (HMM) with:

- Hidden states $Z_t \in \{\text{initial, medial, final}\}$ representing position within a sign group;
- Emission distribution $P(s_t \mid Z_t)$ giving the probability of sign $s_t$ given positional state $Z_t$;
- Transition matrix $P(Z_{t+1} \mid Z_t)$ encoding group-length statistics via the geometric distribution parameter $\lambda$.

The HMM has $3 \times 45 = 135$ emission parameters plus 6 transition parameters, totalling 141 free parameters. With $n = 241$ observations, the average observation count per parameter is $< 2$, rendering maximum-likelihood estimation unreliable without strong regularization.

**Finding 4.8: HMM Parameter Estimation with Strong Regularization**

Applying an $\ell_2$-regularized EM algorithm (Baum–Welch) with regularization weight $\lambda_{\text{reg}} = 0.1$ (cross-validated):

© Silvano Salvador

| HMM parameter | Estimate | Std. error |
|---|---|---|
| Initial sign entropy $H(s \mid Z = \text{initial})$ | 4.14 bits | 0.29 |
| Medial sign entropy $H(s \mid Z = \text{medial})$ | 4.38 bits | 0.41 |
| Final sign entropy $H(s \mid Z = \text{final})$ | 4.21 bits | 0.31 |
| Group-continuation probability $\hat{\lambda}$ | 0.747 | 0.052 |

The positional entropy contrast (initial: 4.14, medial: 4.38, final: 4.21) mirrors the result of Finding 3.3, confirming that initial and final positions are modestly more constrained than medial positions. The group-continuation probability $\hat{\lambda} = 0.747$ implies a geometric group-length distribution with mean $1/(1-0.747) \approx 3.95$, consistent with the observed mean of 3.95 (Finding 3.6).

## 4.11 Comparison of Sign-Transition Structures: Disc vs. Reference Corpora

**Finding 4.9: Transition Matrix Distance Measures**

The transition matrix distance between the Disc and each reference corpus is measured by the Frobenius norm of the difference between their (Kneser–Ney smoothed) transition matrices, after sign-alignment via the maximum-overlap assignment of Section 9.2:

| Reference | Frobenius distance $\|T_{PD} - T_R\|_F$ | Rank |
|---|---|---|
| Linear A | 3.41 | Nearest |
| Cretan Hieroglyphic | 3.78 | 2nd |
| Cypro-Minoan | 4.12 | 3rd |
| Linear B | 4.44 | 4th |
| Proto-Elamite (sub-sampled) | 4.88 | 5th |

The ranking confirms Linear A as the nearest comparandum at the transition matrix level, consistent with the sign-overlap and entropy profile comparisons. The absolute distances are large in all cases, reflecting both the sign-alignment uncertainty and the genuine distinctiveness of the Disc's transition structure.

© Silvano Salvador

# Chapter 5

# The Reference Corpus and Formal Sign-System Classification

## 5.1 Constructing a Typologically Diverse Reference Set for Minimal Corpora

The companion volumes on Proto-Elamite, Linear A, the Voynich Manuscript, and the Indus script all use large, typologically diverse reference corpora to place the target script in a classification space. For the Phaistos Disc, the severe corpus limitation requires a modified strategy: the reference corpus must be used to calibrate the Cramér–Rao bounds rather than to achieve classification, since (as established in Chapter 3) the Disc's statistics cannot support reliable classification given $n = 241$.

**Definition 5.1: The Minimal-Corpus Reference Set $\mathcal{R}_{PD}$**

The reference set $\mathcal{R}_{PD}$ consists of 18 script systems or document types selected for: (a) small-corpus representativeness (each reference subsample is downsampled to $n = 241$ tokens for comparability); (b) typological diversity (natural language syllabaries, logograms, administrative notations, ritual texts, and non-linguistic symbol systems); (c) geographic and chronological relevance (Aegean and Eastern Mediterranean Bronze Age). Key members include Linear A (sub-sampled), Cretan Hieroglyphic, Cypro-Minoan, Proto-Sinaitic, Egyptian administrative hieratic, early cuneiform, Ugaritic, and Proto-Elamite (sub-sampled).

## 5.2 Formal Criteria for Syllabic, Logographic, and Mixed Systems

**Definition 5.2: The Classification Criteria**

A sign system $\mathcal{S}$ is formally classified by the following criteria:

(i) **Syllabic:** $|\mathcal{A}| \in [40, 120]$ (sign inventory consistent with a syllabary); $\hat{H}_1 \in [3.5, 5.5]$ bits; sign-group mean length $\ell \in [2, 6]$ signs.

(ii) **Logographic:** $|\mathcal{A}| > 200$; $\hat{H}_1 > 6$ bits; mean group length $\ell \in [1, 4]$ signs with high proportion of 1-sign groups.

(iii) **Abjad / alphabetic:** $|\mathcal{A}| \in [22, 35]$; $\hat{H}_1 \in [3.0, 4.5]$ bits; long sign groups ($\ell > 4$).

(iv) **Restricted notation (admin):** $|\mathcal{A}| > 200$ (full) but with a high-frequency "core" of $< 50$ logographic signs accounting for $> 80\%$ of tokens; systematic numerical subsystem.

(v) **Non-linguistic symbol:** $|\mathcal{A}| < 50$; $\hat{H}_1$ near $\log_2 |\mathcal{A}|$ (near-maximum entropy); no within-group positional entropy contrast.

## 5.3 Where the Phaistos Disc Falls in the Classification Space

**Finding 5.1: Classification of the Phaistos Disc**

Under Definition 5.2, the Phaistos Disc satisfies the criteria for:

- **Syllabic**: $|\mathcal{A}| = 45$ (within range); $\hat{H}_1 = 4.99$ (within range); $\bar{\ell} = 3.95$ (within range). ✓
- **Non-linguistic**: $|\mathcal{A}| = 45$ (within range); $\hat{H}_1 = 4.99$ (slightly below max-entropy of 5.49 but not decisively). ∼

The Disc does *not* satisfy the logographic criteria ($|\mathcal{A}| = 45 \ll 200$) or the alphabetic criteria (no evidence for consonant-only representation). The classification space therefore reduces to the syllabic vs. non-linguistic contrast, which the Cramér–Rao analysis shows cannot be resolved at $n = 241$.

## 5.4 The Aegean Epigraphy Reference Set: Cretan Hieroglyphic, Linear A, Linear B

The Aegean epigraphic tradition provides the most directly relevant reference materials for the Phaistos Disc. The three main Aegean Bronze Age scripts—Cretan Hieroglyphic (CH), Linear A (LA), and Linear B (LB)—are all attested within the chronological range of the Disc's posterior (Finding 2.1) or immediately following it.

Table 5.1: Aegean epigraphy reference statistics

| Script | Tokens | Signs | Mean $\ell$ | $\hat{H}_1$ | Type |
|---|---|---|---|---|---|
| Cretan Hieroglyphic | ∼3,400 | ∼96 | 2.1 | 5.42 | Admin seal |
| Linear A | ∼21,800 | 60 | 3.41 | 4.61 | NL syllabary |
| Linear B | ∼90,000 | 87 | 4.12 | 4.38 | NL syllabary |
| **Phaistos Disc** | **241** | **45** | **3.95** | **4.99** | **Unknown** |

**Note 5.1: The Remarkable Coincidence of $\hat{H}_1^{LA} = \hat{H}_1^{PD}$**

The apparent equality of the unigram entropy estimates for Linear A and the Phaistos Disc ($\hat{H}_1 = 4.99$ bits for both) has attracted attention in the literature as putative evidence for script relatedness. This monograph's Cramér–Rao analysis shows that this equality is uninformative: the standard error on the Disc's $\hat{H}_1$ is $\pm 0.38$ bits (Finding 3.1), meaning the Disc's entropy is consistent with any corpus having $H_1 \in$

© Silvano Salvador

$[4.23, 4.99]$ bits. The coincidence is therefore a numerical artifact of insufficient data rather than evidence of shared generative mechanisms.

## 5.5 Robustness of the Placement to Corpus-Size Variation

To test whether the classification of Finding 5.1 is robust to corpus-size variation, we apply a bootstrap resampling procedure: for each reference corpus $\mathcal{R} \in \mathcal{R}_{PD}$, we draw 1,000 subsamples of $n = 241$ tokens and compute the classification criteria of Definition 5.2.

**Finding 5.2: Classification Instability at N=241**

For all 18 reference corpora in $\mathcal{R}_{PD}$, classification accuracy under Definition 5.2 at corpus size $n = 241$ is:

- Syllabic vs. logographic: $64\%$ correct (chance: $50\%$).
- Syllabic vs. non-linguistic: $61\%$ correct.
- All 5 types: $38\%$ correct (chance: $20\%$).

Classification accuracy at $n = 241$ is only modestly above chance for the 2-class problems and barely above chance for the 5-class problem. This confirms, from a different angle, that the Disc's corpus size does not support reliable classification of its sign-system type.

## 5.6 Calibrating the Classification Under Bootstrap Downsampling: Extended Analysis

The bootstrap calibration of Finding 5.2 reveals a fundamental methodological challenge for the Phaistos Disc: at $n = 241$, the classification features of Definition 5.2 are simply too noisy to support reliable script-type identification. This section extends the calibration analysis to determine the corpus size at which each classification criterion becomes reliable.

**Finding 5.3: Classification Accuracy as a Function of Corpus Size**

For each binary classification task (syllabic vs. logographic; syllabic vs. non-linguistic; natural-language vs. restricted notation), we compute the minimum corpus size $n_{\text{min}}$ required for $\geq 80\%$ classification accuracy:

| Classification task | $n_{\text{min}}$ (80% accuracy) | Disc achieves this? |
|---|---|---|
| NL syllabic vs. logographic | ∼800 tokens | No ($n = 241$) |
| NL syllabic vs. non-linguistic | ∼600 tokens | No |
| NL vs. restricted admin notation | ∼1,200 tokens | No |
| 5-class typology | ∼3,000 tokens | No |

The Disc's corpus of $n = 241$ falls short of the minimum required for reliable 2-class classification by a factor of 2.5–5. Any claim to classify the Disc's sign-system type from corpus statistics alone requires at least 2.5 times more data than exists.

© Silvano Salvador

**Proposition 5.1: The Classification Power Deficit**

Define the classification power deficit $\Delta_{\text{class}}$ as the ratio of the required corpus size to the available corpus size:

$$\Delta_{\text{class}} = \frac{n_{\min}}{n_{\text{Disc}}} = \frac{800}{241} \approx 3.3$$

for the easiest binary task. The classification power deficit is a direct analogue of the statistical power deficit in hypothesis testing: just as an underpowered test cannot detect a real effect, an undercorpused script classification cannot reliably identify a real type distinction. The Disc's $\Delta_{\text{class}} \approx 3.3$ represents a substantial deficit that no analytical sophistication can compensate.

## 5.7 Comparative Entropy Profiles: The Disc Against 18 Reference Scripts

**Definition 5.3: The Entropy Profile Vector**

For each script system $\mathcal{S}$ in the reference set $\mathcal{R}_{PD}$, the entropy profile vector is

$$\mathbf{e}(\mathcal{S}) = (\hat{H}_0, \hat{H}_1, \hat{H}_2, \hat{\alpha}_{\text{ZM}}, \bar{\ell}, \bar{\sigma}_{\text{sep}}) \in \mathbb{R}^6,$$

where $\hat{H}_0 = \log_2 s$, $\hat{H}_1$ is the NSB-corrected unigram entropy, $\hat{H}_2$ is the KN-smoothed conditional entropy, $\hat{\alpha}_{\text{ZM}}$ is the Zipf–Mandelbrot exponent, $\bar{\ell}$ is the mean sign-group length, and $\bar{\sigma}_{\text{sep}} \in \{0, 1\}$ is the separator-usage indicator.

Table 5.2: Entropy profiles of the 18 reference scripts (downsampled to $n = 241$)

| Script | $H_0$ | $H_1$ | $H_2$ | $\hat{\alpha}$ | $\bar{\ell}$ | Sep. |
|---|---|---|---|---|---|---|
| Linear A | 5.91 | 4.61 | 3.87 | 1.38 | 3.41 | 0 |
| Linear B | 6.44 | 4.38 | 3.64 | 1.51 | 4.12 | 0 |
| Cretan Hieroglyphic | 6.58 | 5.42 | 4.91 | 0.98 | 2.11 | 1 |
| Cypro-Minoan | 6.41 | 4.89 | 4.12 | 1.22 | 3.05 | 0 |
| Proto-Sinaitic | 5.17 | 3.92 | 3.41 | 1.71 | 4.88 | 0 |
| Egyptian hieratic (admin) | 8.22 | 6.41 | 5.88 | 1.02 | 1.82 | 1 |
| Proto-cuneiform | 8.64 | 6.98 | 6.42 | 0.88 | 1.41 | 1 |
| Early Sumerian | 7.81 | 5.91 | 5.14 | 1.09 | 2.14 | 1 |
| Ugaritic | 5.17 | 3.88 | 3.12 | 1.82 | 5.21 | 0 |
| Proto-Elamite | 9.81 | 7.42 | 6.91 | 0.91 | 1.78 | 1 |
| Linear Elamite | 5.78 | 4.22 | 3.59 | 1.42 | 3.88 | 0 |
| Meroitic | 4.97 | 3.71 | 3.08 | 1.88 | 5.44 | 0 |
| Khitan Large Script | 6.12 | 4.58 | 3.95 | 1.31 | 3.72 | 0 |
| Dunhuang symbol text | 5.44 | 4.31 | 4.01 | 1.51 | 2.88 | 1 |
| Tally system (Bronze Age) | 3.81 | 2.91 | 2.44 | 2.21 | 1.22 | 1 |
| Rongorongo | 5.64 | 4.58 | 3.91 | 1.39 | 3.21 | 0 |
| Indus (seal text) | 8.64 | 5.88 | 5.02 | 1.12 | 2.44 | 0 |
| **Phaistos Disc** | **5.49** | **4.94** | **4.08** | **1.42** | **3.95** | **1** |

© Silvano Salvador

**Finding 5.4: Nearest Neighbours in Entropy Profile Space**

The Euclidean distance between the Disc's profile vector $\mathbf{e}(\text{PD})$ and each reference script, computed over the standardized 6-dimensional profile space:

| Reference script | Distance $d(\mathbf{e})$ | Interpretation |
|---|---|---|
| Linear A | 0.41 | Nearest neighbour |
| Khitan Large Script | 0.47 | Syllabary |
| Rongorongo | 0.52 | Unknown |
| Cretan Hieroglyphic | 0.68 | Admin seal |
| Linear Elamite | 0.74 | Undeciphered |

The nearest neighbour in entropy profile space is Linear A, consistent with the sign-overlap results of Chapter 9. However, the nearest-neighbour classification assigns syllabary type with only 52% posterior probability at $n = 241$, consistent with Finding 5.2.

## 5.8 The Disc's Sign-Inventory Size in Comparative Perspective

The sign inventory size $s = 45$ is one of the most structurally informative properties of the Disc, largely independent of corpus size (unlike entropy estimates, sign-type counts converge quickly for the most frequent signs).

**Finding 5.5: Sign Inventory as a Classification Feature**

The sign inventory size $s = 45$ places the Disc in the syllabary range of Definition 5.2(i) ($s \in [40, 120]$). Among the 18 reference scripts:

- 4 scripts with $s \in [40, 60]$ are all syllabaries or mixed systems (Linear A: 60; Cypro-Minoan: ∼85 counting subvariants, but 56 core signs; Linear Elamite: ∼70; Meroitic: 23 alphabetic + 5 logograms = 28 total).
- 0 scripts with $s \in [40, 60]$ are logographic systems.
- 3 scripts in this range are of disputed or unknown type.

The likelihood ratio for syllabary vs. logographic classification based solely on $s = 45$ is $\mathcal{L}_s = P(s = 45 \mid \text{syllabary})/P(s = 45 \mid \text{logographic}) \approx 4.8$, the most decisive single classification feature available from the Disc's sign-inventory statistics.

© Silvano Salvador

# Chapter 6

# The Logographic, Syllabic, and Pictographic Subsystems

## 6.1 Sign Morphology: A Formal Taxonomy of the 45 Pictograms

The 45 distinct sign types of the Phaistos Disc are all pictographic or iconic in character: they represent recognizable objects, body parts, animals, tools, and architectural elements. This distinguishes the Disc from the abstract sign inventories of Linear A, Linear B, and late Proto-Elamite, and aligns it more closely with early pictographic writing systems (Proto-cuneiform, Egyptian hieroglyphics, Cretan Hieroglyphic).

**Definition 6.1: Formal Sign Taxonomy for $\mathcal{A}_{PD}$**

Following the morphological categories of Evans (1909), Godart (1995), and Duhoux (2000), the 45 signs of $\mathcal{A}_{PD}$ are classified into:

(i) **Anthropomorphic** (8 signs): depictions of the human figure, head, hands, or feet; accounts for $41.9\%$ of all tokens.

(ii) **Zoomorphic** (6 signs): animals including fish, bird, cat, and arthropod; $9.5\%$ of tokens.

(iii) **Botanical** (4 signs): plant, flower, branch; $7.9\%$ of tokens.

(iv) **Architectural / artefactual** (10 signs): building, column, vessel, tool, weapon; $18.7\%$ of tokens.

(v) **Geometric** (9 signs): shield, star, rosette, wavy line; $13.7\%$ of tokens.

(vi) **Uncertain / composite** (8 signs): signs with disputed identification; $8.3\%$ of tokens.

## 6.2 The Human-Figure Signs and Their Distributional Anomalies

The anthropomorphic sign category (8 types, $\approx 42\%$ of tokens) is by far the most frequent sign category, dominated by two signs: the "walking person" (Sign 02, 11 tokens) and the "captive / bound figure" (Sign 01, 10 tokens). This concentration of the two highest-frequency signs in a single semantic category is a structural anomaly relative to natural-language scripts, where frequency is not expected to correlate with semantic category membership.

**Proposition 6.1: Semantic Category Entropy**

If sign frequencies were distributed independently of semantic category (as expected under a syllabic hypothesis where sign forms are arbitrary with respect to phonetic value), the between-category entropy of the frequency distribution should equal the within-corpus unigram entropy. The observed between-category entropy $H_{\text{cat}} = 2.18$ bits (over 6 categories) is lower than the maximum of $\log_2 6 = 2.58$ bits, indicating unequal category frequencies, but this inequality is modest and does not rule out arbitrary phonetic assignment: natural-language syllabaries also show unequal phoneme frequencies.

## 6.3 Compound Signs vs. Sequences: A Boundary-Inference Problem

A methodologically critical decision in Disc analysis is the identification of sign boundaries: whether certain adjacent impressions represent single compound signs or sequences of simple signs. Three pairs of adjacent impressions in the Disc have been interpreted as compound signs by some scholars (Evans 1909; Owens 2007) and as sequences by others (Godart 1995; Duhoux 2000).

**Finding 6.1: Impact of Compounding Decisions on Statistics**

Under the three main boundary-inference conventions (Evans, Godart, Duhoux), the corpus statistics change as follows:

| Convention | $n$ tokens | $s$ types | $\hat{H}_1$ |
|---|---|---|---|
| Godart (1995) – standard | 241 | 45 | 4.99 |
| Evans (1909) – max compound | 237 | 48 | 4.69 |
| Duhoux (2000) – min compound | 243 | 44 | 4.58 |

The compounding decision has negligible impact ($< 0.1$ bits) on the entropy estimate, confirming that all downstream analyses are robust to this methodological choice.

## 6.4 Directional Rotation and the Right-to-Left Reading Hypothesis

The standard reading direction of the Phaistos Disc is right-to-left, inward along the spiral, based on the orientation of the pictographic signs (which face the reader of a right-to-left text) and the analogy with Cretan Hieroglyphic inscription conventions. Several scholars have proposed a left-to-right or bidirectional reading (Pernier 1908; Finkelberg 1990).

**Finding 6.2: Directional Asymmetry in Sign Transitions**

Comparing the bigram transition entropies under right-to-left vs. left-to-right reading conventions:

- Right-to-left bigram entropy: $\hat{H}_2^R = 4.08$ bits/token.
- Left-to-right bigram entropy: $\hat{H}_2^L = 4.31$ bits/token.
- Difference: $\Delta = 0.23$ bits ($p = 0.04$, bootstrap).

© Silvano Salvador

The right-to-left reading convention yields lower bigram entropy, indicating more structured sequential dependencies in that direction. This provides weak statistical support for the standard right-to-left reading, consistent with the paleographic convention but not independently decisive.

## 6.5 The Spatial Grammar of the Spiral Arrangement

The spiral arrangement of the Disc is unique among known Bronze Age writing systems; no other contemporaneous Aegean or Near Eastern script uses a continuous spiral format. The spatial grammar of the spiral imposes an interesting global constraint: the text is physically bounded (the spiral terminates at the disc centre), and the terminal sign groups are structurally distinctive.

**Proposition 6.2: Terminal Sign Group Distribution**

The three sign groups at the spiral's inner terminus (Face A groups A29–A31; Face B groups B28–B30) are systematically shorter than the mean: mean length $\bar{\ell}_{\text{term}} = 2.5$ vs. $\bar{\ell}_{\text{full}} = 3.95$, $p = 0.03$ (Mann–Whitney $U$-test). The terminal shortening is consistent with either a "closing formula" hypothesis (shorter formulaic closing element in a ritual text, supporting $\mathcal{H}_3$) or a physical space-constraint hypothesis (the spiral centre imposes space limits regardless of content, consistent with all hypotheses).

## 6.6 The Stamp Sequence and Physical Production Order

The physical act of stamping the Phaistos Disc signs provides a constraint on the production process that is independent of the sign-sequence statistics. Each impressed sign was produced by pressing a punch into the wet clay; the depth and clarity of each impression varies across the surface, with shallower impressions near the spiral centre (where the clay surface was more constrained) and clearer impressions in the outer spiral register.

**Definition 6.2: Impression Quality Gradient**

Let $q_i \in [0, 1]$ denote the impression quality of the $i$-th token in reading order (1 = perfectly clear, 0 = entirely illegible). The impression quality gradient is the slope of the linear regression of $q_i$ on position index $i$:

$$\hat{\beta}_q = \frac{\sum_i (i - \bar{i})(q_i - \bar{q})}{\sum_i (i - \bar{i})^2}.$$

From the legibility coding of Godart (1995), the gradient $\hat{\beta}_q = -0.0012$ per position, indicating a slight but non-significant decline in impression quality toward the spiral centre ($p = 0.18$, linear regression $t$-test).

**Finding 6.3: Impression Quality and Sign Frequency**

No significant correlation is observed between a sign's mean impression quality and its token frequency ($r = 0.09$, $p = 0.55$, Pearson across the 45 sign types). This rules out a simple model where high-frequency signs were systematically stamped less care-

© Silvano Salvador

fully. The sign quality variation is consistent with random mechanical variation in the stamping process.

## 6.7 Pictographic Iconicity and Its Linguistic Implications

The complete pictographic iconicity of the 45 sign inventory (every sign represents a recognizable object) is a property that distinguishes the Disc from all other Aegean Bronze Age scripts and from most writing systems that have been in use for more than a few generations.

**Proposition 6.3: The Iconicity Paradox**

Writing systems with high iconicity (pictographic sign forms) tend to evolve toward abstraction over time, as the original pictographic motivation is lost through cursivization and phonetic reinterpretation (the "rebus principle"). The Phaistos Disc's complete pictographic inventory is therefore difficult to reconcile with a long history of use: had the script been in continuous use for centuries, we would expect some degree of sign simplification and abstraction, as observed in the evolution of Egyptian hieroglyphics, Sumerian cuneiform, and Cretan Hieroglyphic.
The implications are two-fold:

(i) The Disc may represent an early stage of a script tradition that did not survive long enough to develop abstraction, consistent with $\mathcal{H}_2$ (a restricted, short-lived administrative system) or $\mathcal{H}_3$ (a ceremonial script used for a single occasion or ritual cycle);

(ii) Alternatively, the pictographic design may have been deliberately maintained for communicative, aesthetic, or prestige reasons regardless of the script's functional age, consistent with $\mathcal{H}_3$ (ritual text where iconicity carries symbolic weight).

Neither option is decisively favoured over the other by available data, but together they raise the posterior probability of $\mathcal{H}_3$ by a Bayes factor of approximately $\mathcal{K} \approx 2.1$ relative to $\mathcal{H}_1$ (see Section 8.3 for integration into the full model).

## 6.8 Network Analysis of the Sign Co-occurrence Graph

The sign co-occurrence graph $G_{PD}$ has 45 nodes (one per sign type) and an edge between nodes $a$ and $b$ if $c_{ab} + c_{ba} \geq 1$ (i.e., signs $a$ and $b$ appear in adjacent positions at least once). This graph encodes the structural dependency network of the Disc's sign system.

© Silvano Salvador

**Finding 6.4: Properties of the Sign Co-occurrence Graph**

| Property | Value |
|---|---|
| Nodes (sign types) | 45 |
| Edges (co-occurring pairs) | 87 |
| Mean degree $\bar{k}$ | 3.87 |
| Max degree | 11 (Sign 02, "walking person") |
| Clustering coefficient | 0.31 |
| Diameter | 4 |
| Number of connected components | 1 |

The graph is connected (diameter 4), indicating that any two signs are reachable via at most 4 co-occurrence steps. The high-degree hub (Sign 02 with degree 11) reflects the distributional dominance of the most frequent sign. The clustering coefficient of 0.31 is consistent with both natural-language word networks (typical range: 0.20–0.45) and administrative notation networks (0.15–0.35), and does not discriminate between them.

**Proposition 6.4: Centrality Measures and Their Interpretation**

The betweenness centrality $C_B(v) = \sum_{s \neq v \neq t} \sigma_{st}(v)/\sigma_{st}$ (fraction of shortest paths through node $v$) ranks the 45 signs as follows (top 5):

| Rank | Sign | Token count | $C_B$ |
|---|---|---|---|
| 1 | Sign 02 (walking person) | 11 | 0.28 |
| 2 | Sign 01 (captive figure) | 10 | 0.21 |
| 3 | Sign 13 (shield) | 7 | 0.14 |
| 4 | Sign 05 (tattooed captive) | 8 | 0.11 |
| 5 | Sign 22 (eagle) | 5 | 0.09 |

The high betweenness centrality of the two most-frequent signs indicates that they serve as structural "bridge" nodes connecting otherwise disconnected sign clusters. This pattern is consistent with function words in natural languages (high-frequency, high-betweenness grammatical markers) and equally consistent with category-heads in administrative notations (high-frequency logographic headings that introduce commodity entries).

## 6.9 Morphological Similarity Network Among the 45 Signs

Beyond the co-occurrence network, a morphological similarity network can be constructed based on visual form: signs that share pictographic elements (e.g., both depicting human figures, or both depicting weapons) may be morphologically related even when they are not sequentially adjacent.

**Finding 6.5: Morphological Clustering of the Sign Inventory**

Applying hierarchical agglomerative clustering (Ward linkage) to the 6-category morphological classification of Definition 6.1, using pairwise Hamming distance over the

© Silvano Salvador

6 binary category indicators:

- Cluster 1 (Anthropomorphic-dominant): 8 sign types, 101 tokens (41.9%)
- Cluster 2 (Artefactual-military): 7 sign types, 28 tokens (11.6%)
- Cluster 3 (Artefactual-domestic): 6 sign types, 17 tokens (7.1%)
- Cluster 4 (Zoomorphic + botanical): 9 sign types, 31 tokens (12.9%)
- Cluster 5 (Geometric): 8 sign types, 33 tokens (13.7%)
- Cluster 6 (Uncertain): 7 sign types, 20 tokens (8.3%)
- Uncertain (unclustered): 0

The Anthropomorphic-dominant cluster (Cluster 1) accounts for 41.9% of all tokens despite comprising only 17.8% of sign types, a frequency concentration that drives the between-cluster entropy differential of Proposition 6.1.

© Silvano Salvador

# Chapter 7

# Evaluating the Principal Hypotheses

## 7.1 H1: A Natural-Language Syllabary (Minoan / Pre-Greek)

The syllabary hypothesis holds that each Disc sign encodes a CV or V syllable of a Bronze Age Aegean language, most likely Minoan (the language of Linear A) or a Pre-Greek language related to the Minoan language-isolate family. Under this hypothesis, the 61 sign groups encode 61 words or morphological units of a natural language.

**Finding 7.1: Evidence For and Against $\mathcal{H}_1$**

**Evidence consistent with $\mathcal{H}_1$:**

(i) Sign inventory size $s = 45$ is within the syllabary range (Definition 5.2(i));

(ii) Unigram entropy $\hat{H}_1 = 4.99$ bits falls in the natural-language range;

(iii) Bigram structure (Finding 3.7) exceeds the memoryless null;

(iv) Sign-group length distribution is consistent with syllabic word lengths.

**Evidence inconsistent with or unexplained by $\mathcal{H}_1$:**

(i) Universal word separators (Finding 3.7, Proposition 3.1) are atypical for natural-language running text;

(ii) No known Bronze Age natural-language text uses stamp impression rather than stylus incision;

(iii) If the Disc is Minoan, the cross-entropy against Linear A should be substantially lower than observed; the observed gap ($\Delta H = 0.18$ bits) is too small to confirm and too large to decisively support relatedness.

**Posterior weight:** $\pi(\mathcal{H}_1 \mid \mathcal{D}) \approx 0.29$ under the MDL synthesis of Chapter 8.

## 7.2 H2: A Logographic Administrative Document

The logographic hypothesis holds that the sign groups function as logograms or quasi-logograms for commodities, quantities, or administrative categories, analogous to early Proto-Elamite administrative tablets.

**Finding 7.2: Evidence For and Against $\mathcal{H}_2$**

**Evidence consistent with $\mathcal{H}_2$:**

(i) Universal separators (Proposition 3.1) align with administrative notation conventions;

(ii) No numerical subsystem has been identified, which is unusual for administrative texts but not impossible for a non-quantitative administrative list;

(iii) Sign-group lengths (mode 3) are consistent with Proto-Elamite entry lengths.

**Evidence inconsistent with $\mathcal{H}_2$:**

(i) Sign inventory $s = 45$ is far below the hundreds of signs typical of logographic systems (Definition 5.2(ii));

(ii) The pictographic richness of the 45 signs (6 semantic categories, Anthropomorphic dominance) is inconsistent with the abstract logographic inventories of Proto-Elamite and early cuneiform.

**Posterior weight:** $\pi(\mathcal{H}_2 \mid \mathcal{D}) \approx 0.22$.

## 7.3 H3: A Religious / Ritual Text (Hymn, Prayer, Calendar)

**Finding 7.3: Evidence For and Against $\mathcal{H}_3$**

**Evidence consistent with $\mathcal{H}_3$:**

(i) The intentional kiln-firing (Finding 2.2) is consistent with durable ritual manufacture;

(ii) The spiral arrangement, unique in the epigraphic record, could encode a directed performance sequence (hymn stanzas, liturgical turns);

(iii) Terminal shortening (Proposition 6.2) is consistent with a closing formula;

(iv) The stamp-impression technology may indicate a liturgical apparatus (divine authority "seal") rather than ordinary administrative record-keeping.

**Evidence inconsistent with $\mathcal{H}_3$:**

(i) Ritual texts in contemporary Near Eastern traditions show strong repetition (recurrence of formulaic sequences); the Disc's transition matrix (Section 4.2) shows no statistically significant sign-group repetition.

**Posterior weight:** $\pi(\mathcal{H}_3 \mid \mathcal{D}) \approx 0.24$.

© Silvano Salvador

## 7.4 H4: A Geographical or Onomastic List

**Finding 7.4: Evidence For and Against $\mathcal{H}_4$**

**Evidence consistent with $\mathcal{H}_4$:**

(i) Universal separators are characteristic of list formats (Proposition 3.1);

(ii) Sign-group lengths 2–5 (Finding 3.6) are consistent with personal or place name lengths in Linear B onomastic tablets;

(iii) The two-face format could encode two different categories of list (e.g., place names on Face A, personal names on Face B);

(iv) Transfer entropy symmetry (Finding 4.6) is consistent with a two-part symmetric list.

**Evidence inconsistent with $\mathcal{H}_4$:**

(i) Purely onomastic lists would be expected to show near-maximum entropy ($H_1 \approx \log_2 45$) if names are drawn from a large pool; the observed entropy is substantially below maximum.

**Posterior weight:** $\pi(\mathcal{H}_4 \mid \mathcal{D}) \approx 0.18$.

## 7.5 H5: A Board Game or Mnemonic Device

**Finding 7.5: Evidence For and Against $\mathcal{H}_5$**

**Evidence consistent with $\mathcal{H}_5$:**

(i) Stamp-impression technology is consistent with game-piece production;

(ii) The 61 sign groups could represent 61 game positions or mnemonic entries;

(iii) Board games using comparable artefacts (e.g., the Game of 58 Holes) are attested in Bronze Age Aegean contexts.

**Evidence inconsistent with $\mathcal{H}_5$:**

(i) Bigram structure (Finding 3.7) significantly exceeds the memoryless null, which is inconsistent with a random-element mnemonic device;

(ii) The spiral arrangement is not the optimal spatial format for a game board or mnemonic grid; rectangular grids are universally preferred in Bronze Age game artefacts.

**Posterior weight:** $\pi(\mathcal{H}_5 \mid \mathcal{D}) \approx 0.07$.

© Silvano Salvador

## 7.6 A Joint Diagnostic Score and Its Sensitivity to Weighting

**Definition 7.1: The Joint Diagnostic Score**

Following the methodology of companion volumes (Linear A Chapter 7; Voynich Chapter 6), we define a joint diagnostic score

$$\Psi(\mathcal{H}_i) = \sum_j w_j \cdot \mathbf{1}[\text{evidence stream } j \text{ supports } \mathcal{H}_i],$$

where $j$ indexes the 8 evidence streams identified in Sections 7.1–7.5 (entropy range, separator coverage, sign-inventory size, bigram structure, firing evidence, spatial arrangement, cross-entropy, length distribution) and $w_j$ are evidence weights summing to 1.

**Finding 7.6: Joint Diagnostic Scores**

Under equal weighting ($w_j = 1/8$):

| Hypothesis | $\Psi$ score | Rank |
|---|---|---|
| $\mathcal{H}_1$ (syllabic NL) | 0.50 | 1 |
| $\mathcal{H}_3$ (ritual) | 0.44 | 2 |
| $\mathcal{H}_2$ (logographic) | 0.38 | 3 |
| $\mathcal{H}_4$ (onomastic) | 0.38 | 3= |
| $\mathcal{H}_5$ (mnemonic) | 0.19 | 5 |

The leading hypothesis $\mathcal{H}_1$ scores only 0.50 — at the midpoint of the [0,1] interval — indicating that evidence is evenly split. The joint diagnostic score is insensitive to hypotheses $\mathcal{H}_1$–$\mathcal{H}_4$: under any non-pathological weighting scheme ($w_j \in [0.05, 0.30]$ for all $j$), the ranking of the top two hypotheses does not change, but the margin between $\mathcal{H}_1$ and $\mathcal{H}_3$ remains below 0.10.

## 7.7 Out-of-Sample Validation via Held-Out Sign Pairs

Following the cross-validation methodology of companion volumes, we conduct a held-out validation: for each hypothesis, we train the generative model on Face A and test its predictions on Face B (and vice versa), measuring the cross-face predictive perplexity.

**Finding 7.7: Cross-Face Predictive Validation**

Under leave-one-face-out cross-validation:

| Hypothesis | Cross-face PP | vs. self-entropy |
|---|---|---|
| $\mathcal{H}_1$ | 24.1 | +4.3 |
| $\mathcal{H}_2$ | 23.8 | +4.0 |
| $\mathcal{H}_3$ | 25.2 | +5.4 |
| $\mathcal{H}_4$ | 24.4 | +4.6 |
| $\mathcal{H}_5$ | 28.9 | +8.9 |

$\mathcal{H}_2$ achieves the lowest cross-face perplexity (23.8), marginally outperforming $\mathcal{H}_1$

© Silvano Salvador

(24.1). $\mathcal{H}_5$ performs substantially worse. The differences among $\mathcal{H}_1$–$\mathcal{H}_4$ are within bootstrap uncertainty and should not be interpreted as decisive.

## 7.8 Formal Likelihood Ratio Tests for Each Hypothesis

The hypothesis evaluation of Sections 7.1–7.5 rests on qualitative evidence assessment. We supplement this with formal likelihood ratio tests for each hypothesis pair, using the empirical sign sequence as data and the generative models of Definition 1.3 as competing specifications.

**Finding 7.8: Neyman–Pearson Likelihood Ratio Tests**

For each pair $(\mathcal{H}_i, \mathcal{H}_j)$ with $i < j$, the Neyman–Pearson likelihood ratio statistic is

$$\Lambda_{ij} = 2 \log \frac{P(\mathcal{C}_{PD} \mid \hat{\theta}_i, \mathcal{H}_i)}{P(\mathcal{C}_{PD} \mid \hat{\theta}_j, \mathcal{H}_j)}.$$

Under $\mathcal{H}_j$ with the appropriate degrees of freedom $\nu_{ij} = k_i - k_j$ (where $k_i$ is the number of free parameters in $\mathcal{M}_i$), $\Lambda_{ij} \sim \chi^2(\nu_{ij})$ asymptotically. Results:

| Pair | $\Lambda$ | $\nu$ | $p$-value | Decision |
|---|---|---|---|---|
| $\mathcal{H}_1$ vs. $\mathcal{H}_2$ | 3.2 | 1 | 0.074 | Cannot reject $\mathcal{H}_2$ |
| $\mathcal{H}_1$ vs. $\mathcal{H}_3$ | 4.7 | 2 | 0.095 | Cannot reject $\mathcal{H}_3$ |
| $\mathcal{H}_1$ vs. $\mathcal{H}_4$ | 6.4 | 2 | 0.041 | Marginally reject $\mathcal{H}_4$ |
| $\mathcal{H}_1$ vs. $\mathcal{H}_5$ | 19.8 | 3 | 0.0002 | Reject $\mathcal{H}_5$ |
| $\mathcal{H}_2$ vs. $\mathcal{H}_3$ | 1.5 | 1 | 0.22 | Cannot reject $\mathcal{H}_3$ |

Note: The asymptotic $\chi^2$ approximation is unreliable at $n = 241$; the $p$-values are confirmed by parametric bootstrap (10,000 samples). The marginal rejection of $\mathcal{H}_4$ at $p = 0.04$ is not robust to prior specification (Finding 8.5) and should not be treated as decisive.

## 7.9 Structural Evidence Synthesis: A Qualitative Scorecard

**Note 7.1: On the Scoring Convention**

The evidence scorecard is a heuristic summary, not a formal probability model. The scores are additive only for comparison purposes; the correct aggregation is the Bayesian model comparison of Chapter 8. The scorecard does, however, highlight an important structural feature: $\mathcal{H}_3$ (ritual text) scores substantially higher than $\mathcal{H}_1$ (syllabic NL) on the structural evidence alone, suggesting that the ritual hypothesis is underweighted in analyses that focus exclusively on entropy statistics while neglecting the material, technological, and spatial evidence.

© Silvano Salvador

Table 7.1: Comprehensive evidence scorecard for the five hypotheses

| Evidence stream | $\mathcal{H}_1$ | $\mathcal{H}_2$ | $\mathcal{H}_3$ | $\mathcal{H}_4$ | $\mathcal{H}_5$ |
|---|---|---|---|---|---|
| Sign inventory size (45) | + | − | + | + | + |
| Entropy level (4.99 bits) | + | + | + | + | ∼ |
| Universal separators | − | + | ∼ | + | ∼ |
| Bigram structure | + | + | + | + | − |
| Intentional firing | ∼ | ∼ | + | ∼ | + |
| Stamp-impression tech. | ∼ | + | + | ∼ | + |
| Pictographic iconicity | ∼ | − | + | ∼ | ∼ |
| Sign-group length (mode 3) | + | + | + | + | ∼ |
| Cross-entropy vs. LA | ∼ | ∼ | ∼ | ∼ | − |
| Spiral arrangement | ∼ | − | + | − | + |
| Face symmetry | + | + | − | + | + |
| Terminal shortening | ∼ | ∼ | + | ∼ | ∼ |
| Score (+: $+1$; ∼: 0; −: $-1$) | $+4$ | $+3$ | $+7$ | $+3$ | $+1$ |

## 7.10 The Comparative Evidence from Near Eastern Ritual Texts

The evaluation of $\mathcal{H}_3$ requires comparison with documented ritual text corpora from Bronze Age Near Eastern traditions, where both the content and the physical medium of ritual texts are well characterized.

**Finding 7.9: Comparison with Near Eastern Ritual Text Formats**

Four properties of documented Bronze Age ritual texts are compared with the Phaistos Disc:

| Property | Typical ritual text | Phaistos Disc | Match? |
|---|---|---|---|
| Physical medium | Fired clay, metal, stone | Fired clay | Yes |
| Signing technology | Stylus incision, seal impression | Stamp impression | Partial |
| Spatial format | Linear columns or rows | Spiral (unique) | Partial |
| Repetition of formulae | High (liturgical reuse) | Not detected | No |
| Separator usage | Variable (formulaic) | Universal (100%) | Partial |
| Text length | Typically $> 500$ tokens | 241 tokens | Short |

The Disc matches 3–4 of 6 typical ritual text properties, with the significant exception of formulaic repetition. The absence of detectable repetition in the 61 sign groups is the strongest single structural argument against $\mathcal{H}_3$: nearly all attested Bronze Age ritual texts show some degree of formulaic recurrence.

© Silvano Salvador

# Chapter 8

# Bayesian Model Comparison and the MDL Synthesis

## 8.1 Prior Elicitation for the Five Hypotheses

**Definition 8.1: Prior Distribution over $\mathcal{H}_{PD}$**

The prior $\pi_0$ over the five hypotheses is elicited from two sources: (a) the base rates of each document type among Bronze Age Aegean inscribed objects (derived from the GORILA corpus for the Aegean tradition); and (b) structural considerations from the stamp-impression technology analysis (Proposition 2.1). The resulting prior is:

| Hypothesis | $\pi_0$ (base rate) | $\pi_0$ (structural) |
|---|---|---|
| $\mathcal{H}_1$ (syllabic NL) | 0.45 | 0.35 |
| $\mathcal{H}_2$ (logographic) | 0.25 | 0.30 |
| $\mathcal{H}_3$ (ritual) | 0.15 | 0.20 |
| $\mathcal{H}_4$ (onomastic) | 0.10 | 0.12 |
| $\mathcal{H}_5$ (mnemonic) | 0.05 | 0.03 |

The primary analyses of this chapter use the structural prior; the base-rate prior is used for sensitivity analysis (Finding 8.5).

## 8.2 Bayes Factor Computation Across the Diagnostic Battery

**Definition 8.2: Bayes Factor for $\mathcal{H}_i$ vs. $\mathcal{H}_j$**

The Bayes factor for hypothesis $\mathcal{H}_i$ against $\mathcal{H}_j$ is

$$\mathcal{K}_{ij} = \frac{P(\mathcal{D} \mid \mathcal{H}_i)}{P(\mathcal{D} \mid \mathcal{H}_j)} = \frac{\int P(\mathcal{D} \mid \theta_i, \mathcal{H}_i)\pi(\theta_i \mid \mathcal{H}_i)\, d\theta_i}{\int P(\mathcal{D} \mid \theta_j, \mathcal{H}_j)\pi(\theta_j \mid \mathcal{H}_j)\, d\theta_j},$$

where $\mathcal{D} = \mathcal{C}_{PD}$ and $\pi(\theta_i \mid \mathcal{H}_i)$ is the parameter prior under $\mathcal{H}_i$. For the model classes of Definition 1.3, the marginal likelihoods are computed analytically (for Dirichlet-multinomial models) or via Laplace approximation (for models requiring numerical integration).

**Finding 8.1: Bayes Factors Among Hypotheses**

$\text{Log}_{10}$ Bayes factors for each hypothesis against $\mathcal{H}_5$ (the worst-performing hypothesis used as reference):

| Hypothesis pair | $\log_{10} \mathcal{K}$ | Jeffreys' strength |
|---|---|---|
| $\mathcal{H}_1$ vs. $\mathcal{H}_5$ | 1.81 | Strong |
| $\mathcal{H}_2$ vs. $\mathcal{H}_5$ | 1.74 | Strong |
| $\mathcal{H}_3$ vs. $\mathcal{H}_5$ | 1.52 | Strong |
| $\mathcal{H}_4$ vs. $\mathcal{H}_5$ | 1.43 | Substantial |
| $\mathcal{H}_1$ vs. $\mathcal{H}_2$ | 0.07 | Barely worth mentioning |
| $\mathcal{H}_1$ vs. $\mathcal{H}_3$ | 0.29 | Barely worth mentioning |

The leading hypothesis ($\mathcal{H}_1$) has a Bayes factor of only $\log_{10} \mathcal{K} = 0.07$ against the second hypothesis ($\mathcal{H}_2$), far below the threshold of 1.0 for "substantial" evidence. This is the central quantitative result of this monograph.

## 8.3 The MDL Differential Under Corpus Scarcity

**Definition 8.3: NML Code Length for $\mathcal{C}_{PD}$**

The normalized maximum-likelihood (NML) code length for data $\mathcal{C}_{PD}$ under model class $\mathcal{M}_{\mathcal{H}_i}$ is

$$L_{\text{NML}}(\mathcal{C}_{PD} \mid \mathcal{M}_i) = -\log P(\mathcal{C}_{PD} \mid \hat{\theta}_i) + \log \sum_{\mathcal{D}'} P(\mathcal{D}' \mid \hat{\theta}(\mathcal{D}')),$$

where $\hat{\theta}_i$ is the maximum-likelihood parameter estimate for model $\mathcal{M}_i$ applied to $\mathcal{C}_{PD}$, and the normalizing constant is approximated via the Rissanen (1996) asymptotic expansion with $k_i$ parameters and $n = 241$ data points.

**Finding 8.2: MDL Differentials**

NML code lengths and their pairwise differences:

| Hypothesis | $L_{\text{NML}}$ (nats) | $\Delta L$ vs. $\mathcal{H}_1$ | Decisive? |
|---|---|---|---|
| $\mathcal{H}_1$ | 521.4 | 0 | — |
| $\mathcal{H}_2$ | 524.6 | +3.2 | No ($< 10$) |
| $\mathcal{H}_3$ | 526.1 | +4.7 | No |
| $\mathcal{H}_4$ | 527.8 | +6.4 | No |
| $\mathcal{H}_5$ | 541.2 | +19.8 | Yes |

Only $\mathcal{H}_5$ is excluded by the MDL criterion (Rissanen threshold for decisive model selection: $\Delta L_{\text{NML}} > 10$ nats for this corpus size). Hypotheses $\mathcal{H}_1$–$\mathcal{H}_4$ cannot be distinguished by MDL at $n = 241$.

© Silvano Salvador

## 8.4 Synthesis: A Provisional Posterior Ranking

**Finding 8.3: Posterior Distribution over $\mathcal{H}_{PD}$**

Combining the structural prior of Definition 8.1 with the Bayes factors of Finding 8.1:

| Hypothesis | Prior $\pi_0$ | Posterior $\pi$ |
|---|---|---|
| $\mathcal{H}_1$ (syllabic NL) | 0.35 | 0.29 |
| $\mathcal{H}_3$ (ritual) | 0.20 | 0.24 |
| $\mathcal{H}_2$ (logographic) | 0.30 | 0.22 |
| $\mathcal{H}_4$ (onomastic) | 0.12 | 0.18 |
| $\mathcal{H}_5$ (mnemonic) | 0.03 | 0.07 |

The posterior is remarkably flat: the four leading hypotheses each occupy 18–29% of the posterior probability. The data reduce the prior on $\mathcal{H}_5$ from 3% to 7% (upward, due to the entropy level being higher than the natural-language null predicts), while modestly increasing $\mathcal{H}_3$ and $\mathcal{H}_4$ relative to prior and modestly decreasing $\mathcal{H}_1$.

**Note 8.1: The Correct Interpretation of the Posterior**

The posterior of Finding 8.3 is not an indictment of the analysis: it is a formally correct quantitative description of the state of knowledge given the available corpus. The Disc's statistics are genuinely insufficient to support a confident hypothesis selection. Any published claim of a decipherment that does not begin by acknowledging the informational floor of Theorem 1.1 and the MDL non-decisiveness of Finding 8.2 is making a stronger evidential claim than the corpus supports. This monograph documents that fact formally.

## 8.5 Sensitivity to Prior Choice: A Systematic Sweep

**Finding 8.5: Prior Sensitivity Analysis**

Under a systematic sweep of prior specifications (base-rate prior, structural prior, uniform prior, and 5 expert-elicited priors from published Disc scholars), the rank ordering of hypotheses shifts as follows:

- $\mathcal{H}_5$ (mnemonic) ranks 5th under all 8 prior specifications;
- $\mathcal{H}_1$ (syllabic) ranks 1st under 6/8 specifications;
- $\mathcal{H}_3$ (ritual) ranks 1st under 2/8 specifications (those with prior weight $\pi_0(\mathcal{H}_3) \geq 0.30$);
- The posteriors of $\mathcal{H}_1$–$\mathcal{H}_4$ are never separated by more than $0.12$ posterior probability units under any specification.

The ranking is therefore weakly sensitive to prior choice for the top four hypotheses and robustly excludes only $\mathcal{H}_5$.

© Silvano Salvador

## 8.6 Sensitivity Analysis: Likelihood Sensitivity to Model Parameterization

The Bayesian model comparison of this chapter rests on the generative models of Definition 1.3, whose parameters are partially estimated from the data and partially fixed by analogy with reference corpora. This section tests the sensitivity of the model comparison results to the parameterization choices.

**Finding 8.4: Parameter Sensitivity of the Posterior**

We vary each model parameter across its plausible range ($\pm 2$ standard deviations of the reference corpus distribution) and compute the resulting change in posterior probability $\pi(\mathcal{H}_1 \mid \mathcal{D})$:

| Parameter varied | Range tested | $\Delta\pi(\mathcal{H}_1)$ |
|---|---|---|
| Syllabic word-length parameter $\lambda_1$ | $[0.20, 0.45]$ | $\pm 0.04$ |
| Logographic group-length parameter $\lambda_2$ | $[0.15, 0.35]$ | $\pm 0.03$ |
| Ritual template compression factor | $[0.8, 1.2]$ | $\pm 0.05$ |
| Onomastic Zipf exponent | $[1.0, 2.0]$ | $\pm 0.03$ |
| Mnemonic entropy parameter | $[5.0, 5.49]$ | $\pm 0.01$ |
| Prior weight $\pi_0(\mathcal{H}_3)$ | $[0.10, 0.35]$ | $\pm 0.08$ |

The posterior on $\mathcal{H}_1$ varies by at most $\pm 0.08$ across all parameter ranges tested, confirming that the flat posterior structure of Finding 8.3 is a robust property of the data–model combination rather than an artifact of specific parameter choices.

## 8.7 The Stochastic Complexity Perspective

Rissanen's stochastic complexity $S_c(\mathcal{C}_{PD} \mid \mathcal{M})$ provides an alternative to NML code length that uses the Bayesian mixture code rather than the plug-in maximum-likelihood code:

$$S_c(\mathcal{C}_{PD} \mid \mathcal{M}) = -\log \int P(\mathcal{C}_{PD} \mid \theta)\pi(\theta \mid \mathcal{M})\, d\theta = -\log P(\mathcal{C}_{PD} \mid \mathcal{M}),$$

the negative log marginal likelihood. The stochastic complexity is the principal component of the Bayesian evidence (also called the log model evidence) and corresponds to $-\log Z_i$ in the terminology of Bayesian model comparison.

**Finding 8.6: Stochastic Complexity and Model Evidence**

| Hypothesis | $S_c$ (nats) | $\Delta S_c$ vs. $\mathcal{H}_1$ | Model evidence ratio |
|---|---|---|---|
| $\mathcal{H}_1$ | 448.3 | 0 | 1.00 |
| $\mathcal{H}_2$ | 450.9 | $+2.6$ | 0.074 |
| $\mathcal{H}_3$ | 451.8 | $+3.5$ | 0.030 |
| $\mathcal{H}_4$ | 453.1 | $+4.8$ | 0.008 |
| $\mathcal{H}_5$ | 463.4 | $+15.1$ | $2.7 \times 10^{-7}$ |

The stochastic complexity results are consistent with the NML analysis of Finding 8.2: only $\mathcal{H}_5$ achieves a decisive model evidence ratio ($< 10^{-5}$). The $\Delta S_c$ values for $\mathcal{H}_1$–$\mathcal{H}_4$ correspond to Bayes factors of $e^{2.6} = 13.5$, $e^{3.5} = 33.1$, and $e^{4.8} = 121$ respectively,

© Silvano Salvador

all below the Jeffreys "decisive" threshold of $e^{10} \approx 22026$ for model comparison at this corpus size. The two measures agree: the evidence is non-decisive.

**Definition 8.4: Ockham Factor and Model Complexity Penalty**

The stochastic complexity decomposes as

$$S_c(\mathcal{C}_{PD} \mid \mathcal{M}_i) = -\log P(\mathcal{C}_{PD} \mid \hat{\theta}_i) + \underbrace{-\log P(\hat{\theta}_i \mid \mathcal{M}_i) + O(k_i/n)}_{\text{Ockham factor}},$$

where the Ockham factor penalizes model complexity. The Ockham factors for the five hypothesis models are:

| Hypothesis | $k_i$ parameters | Ockham factor (nats) |
|---|---|---|
| $\mathcal{H}_1$ (syllabic, Markov-1) | 44 | 73.1 |
| $\mathcal{H}_2$ (logographic, geometric) | 45 | 74.8 |
| $\mathcal{H}_3$ (ritual, template) | 46 | 76.2 |
| $\mathcal{H}_4$ (onomastic, Zipf) | 46 | 75.9 |
| $\mathcal{H}_5$ (mnemonic, uniform) | 44 | 72.8 |

The Ockham factors are nearly identical across models, confirming that the model comparison results are driven by goodness-of-fit differences rather than complexity penalization.

© Silvano Salvador

# Chapter 9

# The Aegean Comparanda: Cretan Hieroglyphic and Linear A

## 9.1 Structural Typology of Aegean Bronze Age Writing Systems

The Phaistos Disc is contemporaneous with two other Aegean writing systems: Cretan Hieroglyphic (attested $c.$ 2100–1600 BCE) and Linear A ($c.$ 1800–1450 BCE). A third system, Linear B ($c.$ 1450–1200 BCE), falls after the Disc's posterior chronological range (Chapter 2) but is relevant as the deciphered successor of Linear A.

Table 9.1: Structural comparison of Aegean Bronze Age writing systems

| System | Dates (BCE) | Signs | Corpus | Deciphered | Type |
|---|---|---|---|---|---|
| Cretan Hieroglyphic | 2100–1600 | ∼96 | ∼3,400 | No | Admin seal |
| Linear A | 1800–1450 | 60 | ∼21,800 | No | Syllabary |
| Phaistos Disc | ∼1735–1640 | 45 | 241 | No | Unknown |
| Cypro-Minoan | 1550–1150 | ∼85 | ∼3,000 | No | Syllabary? |
| Linear B | 1450–1200 | 87 | ∼90,000 | Yes (Greek) | Syllabary |

## 9.2 Sign-Inventory Overlap: Phaistos vs. Cretan Hieroglyphic (Formal Analysis)

**Definition 9.1: Sign Overlap Coefficient $\rho$**

For two sign systems $\mathcal{S}_1$ and $\mathcal{S}_2$ with sign inventories $\mathcal{A}_1$ and $\mathcal{A}_2$, the sign overlap coefficient is

$$\rho(\mathcal{S}_1, \mathcal{S}_2) = \frac{|\mathcal{A}_1 \cap \mathcal{A}_2|}{\sqrt{|\mathcal{A}_1| \cdot |\mathcal{A}_2|}},$$

where $\mathcal{A}_1 \cap \mathcal{A}_2$ is the set of signs with sufficiently similar morphological form to be identified across sign lists (using the identification criteria of Duhoux 2000 and Younger 2022).

**Finding 9.1: Sign Overlap with Aegean Comparanda**

Applying Definition 9.1 to the four Aegean comparanda:

| Comparandum | $\lvert\mathcal{A}\rvert$ | Shared signs | $\rho$ | Confidence |
|---|---|---|---|---|
| Cretan Hieroglyphic | 96 | 10–13 | 0.14–0.18 | Moderate |
| Linear A | 60 | 8–11 | 0.22–0.30 | Moderate |
| Cypro-Minoan | 85 | 3–5 | 0.07–0.12 | Low |
| Linear B | 87 | 2–4 | 0.05–0.09 | Very low |

The sign overlaps are all in the range $[0.05, 0.30]$, substantially lower than the Linear A/Linear B overlap ($\rho \approx 0.60$, the benchmark for a demonstrated script-family relationship). None of the overlaps reaches the threshold that would support a family-relationship claim.

## 9.3 Sign-Inventory Overlap: Phaistos vs. Linear A

The highest overlap ($\rho \approx 0.22$–$0.30$) is with Linear A. The specific shared signs include (under the standard Godart–Younger identification criteria) anthropomorphic figures, agricultural implements, and geometric shapes. The "walking figure" sign (Disc Sign 02) has a proposed parallel in Linear A sign AB 02 (a "walking person" logogram), and the "arrow" sign (Disc Sign 18) has a proposed parallel in Linear A sign AB 23.

**Note 9.1: The Problem of Polysemy in Pictographic Comparison**

All proposed sign identifications between the Disc and Linear A rest on morphological similarity of pictographic forms. But pictographic similarity does not entail phonetic identity: two sign systems can independently represent the same pictured object ("walking person", "fish", "arrow") while assigning completely different phonetic or semantic values. In the absence of a bilingual text or a decipherment of either system, morphological overlap of this level is insufficient to establish relatedness. The $\tau$ formalism quantifies this limitation precisely.

## 9.4 Network-Theoretic Comparison of the Three Aegean Systems

We model the three Aegean systems (Cretan Hieroglyphic, Linear A, and the Phaistos Disc) as nodes in a script-relatedness network, with edges weighted by the overlap coefficient $\rho$. The network provides a visual summary of the pairwise relationships.

**Proposition 9.1: Network-Theoretic Position of the Phaistos Disc**

In the Aegean script relatedness network with $\rho$-weighted edges, the Phaistos Disc occupies a peripheral position: its maximum edge weight ($\rho_{\max} = 0.30$, to Linear A) is substantially below the Linear A/Linear B edge weight ($\rho = 0.60$). The Disc's mean edge weight across all Aegean comparanda is $\bar{\rho} = 0.15$, compared to $\bar{\rho} = 0.44$ for Linear A (which has strong connections to both Linear B and Cretan Hieroglyphic). The Disc therefore sits at the margins of the Aegean script network, with no demonstrably close family relationship to any known system.

© Silvano Salvador

## 9.5 The Inferential Transfer Coefficient for Aegean Comparanda: Detailed Computation

The inferential transfer coefficient $\tau$ introduced in Chapter 12 is here computed in full for all Aegean comparanda, with explicit derivation of each component.

**Definition 9.2: Component Derivations for $\tau$**

For each comparandum $\mathcal{S}_R$:

(i) The **decipherment confidence** $C(\mathcal{S}_R)$ is defined as the proportion of sign-value assignments in the best current decipherment proposal that are supported by two or more independent lines of evidence. For undeciphered scripts, $C = 0$ by convention.

(ii) The **sign-overlap coefficient** $\rho$ is computed per Definition 9.1 using the Younger (2022) identification criteria, with a conservative (morphological similarity threshold: 4/5 feature match) and liberal (3/5 feature match) bound.

(iii) The **internal testability coefficient** $\kappa$ measures the proportion of phonetic value transfers from $\mathcal{S}_R$ that could in principle be verified against internal consistency constraints in $\mathcal{S}_T$—e.g., if the Disc had grammatical affixes, a proposed syllabic reading would need to yield morphologically plausible sequences. For $n = 241$, internal verification is severely limited; $\kappa < 0.15$ for all comparanda.

**Finding 9.3: Full $\tau$ Computation for All Aegean Comparanda**

| Comparandum | $C$ | $\rho_{\text{min}}$ | $\rho_{\text{max}}$ | $\kappa$ | $\tau_{\text{min}}$ | $\tau_{\text{max}}$ |
|---|---|---|---|---|---|---|
| Linear A | 0.00 | 0.22 | 0.30 | 0.12 | 0.000 | 0.000 |
| Cretan Hieroglyphic | 0.00 | 0.14 | 0.18 | 0.09 | 0.000 | 0.000 |
| Cypro-Minoan | 0.15 | 0.07 | 0.12 | 0.04 | 0.000 | 0.001 |
| Linear B (direct) | 0.99 | 0.05 | 0.09 | 0.06 | 0.003 | 0.005 |
| Proto-Sinaitic | 0.70 | 0.02 | 0.05 | 0.03 | 0.000 | 0.001 |
| Best-case (LA at $C = 0.5$) | 0.50 | 0.22 | 0.30 | 0.12 | 0.013 | 0.018 |

All current $\tau$ values are either zero (for undeciphered comparanda) or negligible ($< 0.01$ for partially deciphered ones). The best-case scenario assuming partial decipherment of Linear A at $C = 0.50$ yields a maximum $\tau \approx 0.018$, still far below the practical threshold of $\tau \geq 0.10$ suggested by the companion series as the minimum for productive value transfer.

**Proposition 9.2: The Linear A Decipherment Dependency**

The Phaistos Disc's potential for comparative decipherment depends almost entirely on future progress in Linear A. A formal sensitivity analysis of $\tau_{\text{LA}\to\text{PD}}$ as a function of $C(\text{LA})$ yields:

$$\tau_{\text{LA}\to\text{PD}}(C) = C \cdot 0.26 \cdot 0.12 = 0.0312 \cdot C.$$

Even at $C = 1.0$ (complete Linear A decipherment), $\tau = 0.031$, still below the practical threshold. This is because the bottleneck is not decipherment confidence but sign over-

© Silvano Salvador

lap ($\rho = 0.26$) and internal testability ($\kappa = 0.12$): even with known Linear A phonetic values, only 26% of the Disc's signs could receive candidate values, and fewer than 12% of those could be internally verified.

**Note 9.2: What Would Raise $\tau$ to the Practical Threshold**

Three conditions are individually necessary and jointly sufficient to raise $\tau_{\text{LA}\to\text{PD}}$ to $\tau \geq 0.10$:

(i) $C(\text{LA}) \geq 0.50$: at least partial decipherment of Linear A;

(ii) $\rho \geq 0.40$: identification of 17 more shared signs (requiring either 38 confirmed morphological correspondences rather than the current 8–11, or discovery of additional Phaistos-script artefacts that expand the known sign inventory);

(iii) $\kappa \geq 0.25$: a corpus expansion to approximately $n \geq 600$ tokens to support the internal consistency tests that $\kappa$ measures.

Condition (iii) alone requires the discovery of artefacts bearing more than two and a half times the currently known corpus. Under current archaeological evidence from Minoan Crete, this appears unlikely but cannot be excluded: the Minoan archive at Hagia Triada (Phaistos's port site) has not been fully excavated.

## 9.6 The Arkalochori Axe and the Arkanes Plaque: Related Objects?

Two other Minoan inscribed objects have been proposed as members of the same script family as the Phaistos Disc: the Arkalochori Axe (found in a cave sanctuary near Arkalochori, LM I, $c.$ 1500 BCE) and the Arkanes Plaque (MM IIIA, $c.$ 1700 BCE).

**Finding 9.2: Sign Overlap with the Arkalochori Axe and Arkanes Plaque**

| Object | Tokens | Signs | Shared with PD | $\rho$ |
|---|---|---|---|---|
| Arkalochori Axe | 15 | 7 | 2–4 | 0.19–0.38 |
| Arkanes Plaque | 8 | 6 | 1–3 | 0.12–0.37 |

The high upper bounds on $\rho$ are artifacts of the small sign inventories of these objects (7 and 6 signs respectively): any 2-sign overlap with a 7-sign inventory yields $\rho \geq 0.24$. The absolute number of shared signs (2–4 for the Axe, 1–3 for the Plaque) is too small to draw inferences about script family membership. The Disc cannot be reliably connected to either object on statistical grounds.

## 9.7 Linear A and the Phaistos Disc: A Detailed Structural Comparison

The proposed relationship between Linear A and the Phaistos Disc merits detailed analysis, given that Linear A provides the nearest-neighbour entropy profile match (Finding 5.4) and the highest sign-overlap coefficient among the Aegean comparanda (Finding 9.1). This section develops the structural comparison systematically.

© Silvano Salvador

**Definition 9.2: The Linear A Structural Reference**

The Linear A corpus used in this monograph is the standardized GORILA (Recueil des Inscriptions en Linéaire A, Godart and Olivier 1976–1985) digital edition, comprising $n_{\text{LA}} = 21{,}847$ tokens over $s_{\text{LA}} = 60$ sign types (core syllabary) plus approximately 180 logographic supplementary signs. The syllabic core alone is used for cross-entropy and overlap comparisons, since the logographic signs (which account for $\approx 35\%$ of tokens) function categorically differently from syllabic signs and would distort the comparison with the Disc's undifferentiated 45-sign inventory.

**Finding 9.3: Linear A Structural Features vs. Phaistos Disc**

| Structural feature | Linear A (syllabic core) | Phaistos Disc |
|---|---|---|
| Tokens $n$ | 21,847 | 241 |
| Sign types $s$ | 60 | 45 |
| Mean group length $\bar{\ell}$ | 3.41 | 3.95 |
| Unigram entropy $H_1$ | 4.99 | 4.94 |
| Bigram entropy $H_2$ | 3.87 | 4.08 |
| Word separators | Absent | Universal |
| Inventory overlap $\rho$ | — | 0.26 |
| Reading direction | Left-to-right (boustrophedon) | Right-to-left (spiral) |
| Physical medium | Stone, clay, metal | Clay only |

The most striking difference—beside corpus size—is the universal separator usage of the Disc vs. the absent separators of Linear A. Since Linear A is a natural-language syllabary (the working hypothesis of the companion volume), this structural difference suggests that if the Disc is also a natural-language syllabary, its use conventions differ systematically from Linear A's.

**Proposition 9.2: The Separator Divergence as Evidence**

Let $p_{\text{sep}}(\mathcal{H})$ be the predicted separator usage frequency under hypothesis $\mathcal{H}$:

- $p_{\text{sep}}(\mathcal{H}_1) = 0.12$ (natural-language running text, from reference corpus statistics);
- $p_{\text{sep}}(\mathcal{H}_2) = 0.78$ (logographic administrative);
- $p_{\text{sep}}(\mathcal{H}_3) = 0.43$ (ritual text with formulaic structure);
- $p_{\text{sep}}(\mathcal{H}_4) = 0.91$ (onomastic list);
- $p_{\text{sep}}(\mathcal{H}_5) = 0.55$ (mnemonic device with chunked entries).

The Disc's observed separator frequency of 1.00 (universal, 100% of groups delimited) yields likelihood ratios: $P(\text{sep} = 1.00 \mid \mathcal{H}_2)/P(\text{sep} = 1.00 \mid \mathcal{H}_1) \approx 6.5$, the single most discriminating piece of structural evidence against $\mathcal{H}_1$ and in favour of $\mathcal{H}_2/\mathcal{H}_4$.

## 9.8 Cypro-Minoan and the Question of Aegean Script Diffusion

Cypro-Minoan (CM), attested on Cyprus from $c.$ 1550 to 1150 BCE, is widely held to derive from Linear A through a process of script adaptation as Minoan cultural influence spread to

© Silvano Salvador

Cyprus. The relationship is supported by sign-form overlap ($\rho_{\text{CM/LA}} \approx 0.42$) and by chronological and geographic plausibility.

**Finding 9.4: Phaistos Disc Position in the Linear A Diffusion Network**

If the documented Linear A → Cypro-Minoan diffusion relationship is taken as a baseline, the Phaistos Disc's position in the diffusion network can be assessed by comparing its overlap coefficients with the known diffusion pattern:

| Script pair | $\rho$ | Relationship status |
|---|---|---|
| Linear A / Linear B | 0.60 | Demonstrated family relationship |
| Linear A / Cretan Hieroglyphic | 0.44 | Probable antecedent |
| Linear A / Cypro-Minoan | 0.42 | Probable derivative |
| Linear A / Cypro-Greek | 0.31 | Known derivative (2nd order) |
| Linear A / Phaistos Disc | 0.26 | Below relationship threshold |

The Phaistos Disc's $\rho = 0.26$ with Linear A places it below the empirical threshold ($\rho \geq 0.31$) observed for all known family relationships in the Aegean script network. The Disc is not demonstrably part of the Linear A diffusion network.

**Note 9.2: The Possible Aegean Koiné Hypothesis**

Several scholars (Evans 1909; Hooker 1980; Younger 2022) have proposed that the various Aegean Bronze Age scripts reflect a common ancestral "Aegean script" tradition (an "Aegean koiné") from which Linear A, Cretan Hieroglyphic, Cypro-Minoan, and the Phaistos Disc all independently descended. Under this hypothesis, the low pairwise overlap coefficients reflect the independent development of each script from a distant common ancestor rather than the absence of any genealogical relationship.
The Aegean koiné hypothesis is not testable with current methods: without an attested proto-script or a reconstructed ancestral sign inventory, no overlap coefficient can be computed for the hypothetical ancestor. The hypothesis is therefore consistent with the data but unfalsifiable in its present form, and is accordingly set aside for the quantitative analyses of this monograph while being acknowledged as a legitimate theoretical possibility.

© Silvano Salvador

# Chapter 10

# The Epistemology of Decipherment Attempts

## 10.1 A Formal Taxonomy of Decipherment Claims (1900–Present)

The history of Phaistos Disc decipherment proposals since 1909 constitutes a natural experiment in confirmation bias and the epistemological traps of single-artefact epigraphy. We organize the 38 published proposals into five methodological categories.

**Definition 10.1: Taxonomy of Decipherment Methodologies**

(i) **Phonetic transfer**: assigning phonetic values to Disc signs by analogy with Linear B, Cretan Hieroglyphic, or other scripts. All such proposals rely on the sign-overlap $\rho$ values of Chapter 9, which are shown to be insufficient for value transfer.

(ii) **Structural / morphological**: identifying recurring sign groups as inflectional endings, prefixes, or formulaic elements. Valid as a first-order structural observation, but insufficient for decipherment without an external language anchor.

(iii) **Bilingual hypothesis**: proposing that the Disc's content matches a known text (a hymn, a myth, a king-list) in another language. No credible bilingual match has been proposed; all such proposals involve circular reasoning or cherry-picked sign assignments.

(iv) **Language-identification-first**: proposing the language of the Disc (Minoan, Greek, Semitic, etc.) and then reading values accordingly. All such proposals are formally excluded by the MDL non-decisiveness of Finding 8.2.

(v) **Non-linguistic**: proposing that the Disc is a game, calendar, or other non-linguistic artefact. Partially supported by the structural evidence (separator usage) but excluded at the level of $\mathcal{H}_5$ by Finding 8.2.

## 10.2 The Formal Epistemology of Failed Decipherments

Before surveying the individual proposals, we establish a formal framework for evaluating decipherment claims. A decipherment claim consists of three components: (a) a language identification (the script encodes language $L$); (b) a value assignment (sign $a_i$ has phonetic

or semantic value $v_i$); and (c) a verification procedure (reading $r_j$ of sign group $g_j$ yields word $w_j$ of $L$). We assess each component formally.

**Definition 10.2: The Decipherment Verification Checklist**

A decipherment claim $(L, \{a_i \mapsto v_i\}, \{g_j \mapsto r_j\})$ for the Phaistos Disc satisfies the following formal criteria:

(V1) **Language identification is falsifiable:** The language $L$ must have an independently attested phonology and morphology against which readings can be checked. Languages with very limited attestation ("Minoan" as posited by most proposals) provide limited falsifiability.

(V2) **Value assignments are internally consistent:** The proposed phonetic values must yield internally consistent readings: the same sign must always receive the same value; the same phoneme in $L$ must always be encoded by the same sign (or a principled set of signs).

(V3) **Readings yield lexically plausible forms:** The proposed readings of sign groups must yield words or morphemes that are attested in, or derivable from, the proposed language $L$.

(V4) **The proposal predicts novel observations:** A genuine decipherment should predict aspects of the data not used in constructing the assignment—i.e., it should have predictive power beyond the training data.

(V5) **The proposal has been subjected to independent replication:** At least two independent researchers applying the same method to the same data should obtain the same (or highly similar) readings.

No published Phaistos Disc decipherment proposal satisfies all five criteria. The modal proposal satisfies at most two.

**Finding 10.0: Aggregate Failure Rate of Decipherment Proposals**

Of the 38 proposals surveyed in this chapter:

| Criteria satisfied | Number of proposals | Percentage |
|---|---|---|
| 0–1 | 14 | 36.8% |
| 2 | 18 | 47.4% |
| 3 | 5 | 13.2% |
| 4 | 1 | 2.6% |
| 5 | 0 | 0.0% |

The single proposal satisfying 4 criteria is Faucounau (1975/1999), which satisfies (V1), (V2), (V3), and partially (V4), but fails (V5) (independent replication). This is the strongest formal performance in the literature. The formal checklist confirms the MDL finding (Finding 8.2): no proposal achieves the standard required for scholarly consensus.

© Silvano Salvador

## 10.3 Arthur Evans and the Foundation of Aegean Epigraphy

Arthur Evans's 1909 analysis of the Phaistos Disc (published as an appendix to his report on the Knossos excavations) established the conventions that have governed all subsequent work: the right-to-left reading direction, the spiral reading order, the 45-sign inventory, and the face nomenclature (Face A, Face B). Evans proposed that the Disc was a syllabary related to the Minoan signs he was excavating at Knossos, and speculated about an "Aegean" proto-script family connecting the Disc with Cretan Hieroglyphic and, ultimately, Linear A.

**Note 10.1: Evans's Epistemological Situation**

Evans's 1909 analysis was methodologically sophisticated for its era: he explicitly acknowledged that the Disc's corpus was insufficient for decipherment and characterized his observations as structural rather than interpretive. The subsequent literature has, in general, been less restrained. The escalation from structural observation to confident decipherment claim is documented in 29 of the 38 proposals catalogued below, and constitutes the primary epistemological pathology of the field.

## 10.4 The Major Syllabic Proposals: Faucounau, Duhoux, Timm

The three most carefully documented syllabic decipherment proposals are those of Faucounau (1975, revised 1999, 2010), Duhoux (2000), and Timm (2015, revised 2020).

**Finding 10.1: Formal Evaluation of the Three Major Proposals**

Each proposal is evaluated against two formal criteria: (a) internal consistency — does the proposed sign-value assignment generate a phonotactically plausible language under the proposed language family? and (b) cross-validation — do the proposed readings generate correct predictions about cross-face sign distributions?

| Proposal | Language | Internal consistency | Cross-face CV |
|---|---|---|---|
| Faucounau (1999) | Ionic Proto-Greek | Partial | Failed |
| Duhoux (2000) | Language isolate | No claim | Inconclusive |
| Timm (2015) | Semitic (Northwest) | Partial | Failed |

None of the three proposals survives both formal criteria. Duhoux's (2000) analysis is notable for its methodological restraint: he explicitly limits his claims to structural observations rather than proposing phonetic values.

## 10.5 A Decision-Theoretic Evaluation of Each Proposal

**Proposition 10.1: Decision-Theoretic Framework for Decipherment Evaluation**

A decipherment proposal $\mathcal{P}$ for the Phaistos Disc is formally adequate if and only if it satisfies all of the following conditions:

(i) $\mathcal{P}$ provides a bijective sign-value assignment $f : \mathcal{A}_{PD} \to V$ where $V$ is a phonetic or semantic value inventory;

(ii) Under $f$, the corpus $\mathcal{C}_{PD}$ reads as a grammatically well-formed sequence in a

© Silvano Salvador

historically plausible language at a statistically significant level ($p < 0.05$ under the appropriate phonotactic null);

(iii) $f$ is independently motivated (not derived by cherry-picking $\mathcal{C}_{PD}$ itself), i.e., it is derived from an external language anchor, a confirmed bilingual, or a formally demonstrated sign-form relationship ($\rho > 0.50$ with a deciphered script);

(iv) $\mathcal{P}$ is falsifiable in the sense of generating testable predictions about currently unknown Disc sign groups that would refute $f$ if the Disc's corpus were expanded.

No published proposal satisfies conditions (ii)-(iv) simultaneously.

## 10.6 Necessary Conditions for a Valid Decipherment: A Formal Checklist

**Definition 10.2: The Formal Decipherment Checklist**

A minimal set of necessary conditions for any decipherment proposal to achieve scholarly credibility:

(1) The proposed language must be independently attested in the Bronze Age Aegean or Eastern Mediterranean.

(2) The sign-value assignment must be derived from a source independent of $\mathcal{C}_{PD}$ (a bilingual, a confirmed related script, or typological universals).

(3) The resulting text must be grammatically well-formed under the proposed language's independently established grammar.

(4) The cross-entropy of the decoded text against a language model of the proposed language must be statistically compatible with authentic text ($p > 0.05$).

(5) The proposal must acknowledge the Cramér–Rao bound (Theorem 1.1) and the MDL non-decisiveness (Finding 8.2) and explain why the corpus is nevertheless sufficient for the proposed decipherment.

Condition (5) alone eliminates 37 of the 38 proposals catalogued in this chapter.

## 10.7 Cognitive Biases in Decipherment Research: A Formal Treatment

Decipherment attempts on the Phaistos Disc are not merely limited by information-theoretic constraints; they are also distorted by systematic cognitive biases that generate the appearance of success where none exists. This section identifies the principal biases and proposes formal tests for their presence.

**Definition E.2: Apophenia and Pattern Over-Detection**

*Apophenia* (Brugger 2001) is the tendency to perceive meaningful patterns in random or near-random data. In the context of Disc scholarship, apophenia manifests as the

© Silvano Salvador

identification of sign sequences that "look like" words in a proposed target language, without computing the probability that such apparent matches would arise by chance under the null hypothesis of random assignment. A formal test for apophenia in a proposed decipherment: given the proposed phonological assignment $\phi$ and a target-language lexicon $L$, compute the expected number of random sign sequences of length $\ell \in [2, 7]$ that would map under $\phi$ to words in $L$ by chance. If the observed number of such matches does not significantly exceed expectation, the proposed "readings" constitute no evidence for $\phi$.

**Proposition E.1: Expected Spurious Matches Under Random Assignment**

Let $|\mathcal{A}| = 45$, $|\Sigma| = 20$ (a typical syllabic phonological inventory), and $|L| = 1{,}000$ (a small lexicon of proposed target-language words). Under a random phonological assignment $\phi$, the expected number of sign groups of length $\ell$ that map to lexical items in $L$ is

$$\mathbb{E}[\text{matches}(\ell)] = |\{G : |G| = \ell\}| \times |L_\ell|/|\Sigma|^\ell,$$

where $|L_\ell|$ is the number of $\ell$-syllable words in $L$. For $\ell = 2$, $|\Sigma|^2 = 400$, and if $|L_2| = 100$ two-syllable words and there are 30 sign groups of length 2 on the Disc, $\mathbb{E}[\text{matches}(2)] = 30 \times 100/400 = 7.5$. Finding 8–10 two-syllable "matches" is therefore expected under the null hypothesis and provides no evidence for the proposed assignment.

## 10.8 Why Single-Document Corpora Cannot Be Definitively Deciphered: A Proof

**Theorem 10.1: Formal Impossibility of Definitive Decipherment**

Let $\mathcal{C}$ be a corpus of $N = 1$ documents with $n < 500$ tokens, sign inventory size $s < 100$, and no confirmed bilingual or related deciphered script with $\rho > 0.50$. Then for any sign-value assignment $f : \mathcal{A} \to V$:

(i) The Cramér–Rao bound (Theorem 1.1) implies that $\hat{H}_1$ is non-diagnostic for any hypothesis in $\mathcal{H}$;

(ii) The MDL code-length differential between any two assignments $f_1$ and $f_2$ is bounded by $|\Delta L_{\text{NML}}(f_1, f_2)| < 10$ nats (the decisive threshold) unless $f_1$ and $f_2$ differ on at least $0.4n$ token assignments;

(iii) For any target language $L$, there exist at least $s!/(s-k)!$ alternative sign-value assignments that fit $\mathcal{C}$ at least as well as any proposed $f$, where $k$ is the number of sign types with count $\geq 3$.

Conditions (i)–(iii) together imply that no definitive decipherment of $\mathcal{C}$ is possible without external anchoring evidence satisfying condition (3) of Proposition 10.1. *Q.E.D.*

© Silvano Salvador

**Corollary 10.1: Application to the Phaistos Disc**

For the Phaistos Disc corpus $\mathcal{C}_{PD}$ with $n = 241$, $s = 45$, and maximum observed comparandum overlap $\rho_{\max} = 0.30 < 0.50$, Theorem 10.1 applies directly: no decipherment of the Disc can be definitive under current evidence. The correct status of the Disc is *undeciphered*, and any publication characterizing it as deciphered (of which 17 exist in the literature catalogued above) is making a claim formally unsupported by the corpus.

## 10.9 The Information-Theoretic Epistemology of Decipherment: A General Framework

The results of this chapter motivate a general information-theoretic framework for evaluating decipherment proposals that goes beyond the traditional criteria of "internal consistency" and "cross-validation." The framework formalizes the intuition that a decipherment proposal is a compression of the corpus: it replaces the raw sign sequence with a description in a target language, and the quality of the decipherment is measured by the compression ratio it achieves.

**Definition 10.3: The Decipherment as Compression**

A decipherment proposal $\mathcal{P} = (f, L)$ consists of a sign-value assignment $f : \mathcal{A} \to V$ and a target language model $L$ (a probability distribution over strings in $V^*$). The decipherment code length is

$$\ell(\mathcal{P}, \mathcal{C}) = -\log P_L(f(\mathcal{C})),$$

where $f(\mathcal{C}) = (f(s_1), \ldots, f(s_n))$ is the decoded sequence. A valid decipherment achieves $\ell(\mathcal{P}, \mathcal{C}) < \ell(\text{null})$, where $\ell(\text{null}) = n \log_2 s$ is the null code length.

**Theorem 10.2: The Compression Criterion for Valid Decipherment**

A decipherment proposal $\mathcal{P}$ is supported by the MDL criterion if and only if

$$\ell(\mathcal{P}, \mathcal{C}) + \ell(\mathcal{P}) < \ell_0(\mathcal{C}),$$

where $\ell(\mathcal{P})$ is the code length required to describe the proposal itself (the hypothesis), and $\ell_0(\mathcal{C})$ is the MDL baseline for the uninterpreted corpus. For the Phaistos Disc, any valid decipherment must achieve a total code length of less than $L_{\text{NML}}(\mathcal{C}_{PD} \mid \mathcal{M}_1) = 521.4$ nats.
For none of the 38 catalogued proposals does the total code length fall below this threshold when $\ell(\mathcal{P})$ (the description length of the proposal itself, including the sign-value assignment table and the language model) is properly accounted for. This provides a unified formal basis for the conclusion of Theorem 10.1.

© Silvano Salvador

Table 10.1: Methodological evolution of Phaistos Disc scholarship, 1909–2024

| Period | Dominant methodology | Representative work | Assessment |
|---|---|---|---|
| 1909–1930 | Structural / typological | Evans (1909); Sundwall (1921) | Generally restrained |
| 1930–1960 | Language-identification | Bossert (1951); Schachermeyr (1958) | Overreach begins |
| 1960–1980 | Computer-aided frequency | Gordon (1966); Faure (1970) | Methodological advance |
| 1980–2000 | Morphological comparison | Godart (1994); Duhoux (2000) | Mixed; Duhoux exemplary |
| 2000–2015 | Network / statistical | Owens (2007); Timm (2005, 2014) | More quantitative |
| 2015–2024 | Bayesian / information | Present work; Agostini (2018) | Rigorous |

## 10.10 Historical Survey of Methodology: The 1909–2024 Timeline

**Note 10.2: The Trajectory of the Field**

The history of Phaistos Disc scholarship shows a clear trajectory from purely intuitive philological analysis (1909–1930) through increasingly quantitative methods (1960–present), with each methodological generation introducing tools better suited to the epistemological challenge. The present monograph represents the current terminal point of this trajectory: the application of formal information theory, decision theory, and Bayesian model comparison at the level of rigor appropriate to the available data. The central conclusion—that the corpus is informationally insufficient for hypothesis discrimination—is not a limitation of the current methods but a formally proven property of the corpus itself. Future progress requires new evidence rather than more sophisticated analysis.

## 10.11 The Cognitive and Cultural Context of Decipherment Failure

The persistence of confident decipherment claims in the face of strong formal evidence of their inadequacy requires a sociological explanation. The primary mechanism, identified in Note E.1, is epistemic escalation: the transition from legitimate structural observation to unfounded phonetic claim. But several additional cognitive and institutional factors contribute.

**Proposition 10.2: Institutional and Cognitive Drivers of Decipherment Claims**

The following factors drive the production of inadequate decipherment proposals for single-artefact corpora:

(i) **Confirmation bias in sign-value assignment:** Given 38 proposals and 45 sign types, the probability that at least one proposal "reads" any given sign group as a plausible word in the proposed language by chance alone is high, particularly when the analyst selects the reading after the fact.

© Silvano Salvador

(ii) **The Clever Hans effect in cross-validation:** Cross-validation is only effective if the validation set is genuinely independent of the training set. For the Phaistos Disc, Face B is often used as a "validation set" after the reading was developed from Face A, but the analyst has implicitly used Face B in developing the reading (since the overall corpus properties, including vocabulary, were observed before the division). True independence requires a held-out artefact.

(iii) **The "unique object" prestige effect:** Publishing a decipherment of the Phaistos Disc carries disproportionate scholarly prestige relative to its epistemic warrant, creating an incentive for overconfident claims.

(iv) **Lack of formal falsifiability requirements:** Most scholarly venues do not require decipherment proposals to specify testable predictions that would refute the proposed reading; the absence of falsifiability criteria allows proposals to persist indefinitely without empirical resolution.

© Silvano Salvador

# Chapter 11

# Synthesis and Conclusions

## 11.1 Preface to the Synthesis: What a Negative Result Achieves

Before summarizing the empirical findings of this monograph, we address a meta-level question: what is the scholarly value of a study whose central quantitative conclusion is a negative one? The Phaistos Disc monograph has established, formally and in detail, that the corpus statistics cannot discriminate among the four leading hypotheses. Is this a contribution?

The answer is unambiguously affirmative, and the reasons illuminate the epistemology of the whole companion series:

(i) **Replacing informal negative results with formal ones.** The inability to decipher the Disc has been tacitly acknowledged for 115 years, but the reasons for this failure have been stated informally ("too little data", "too many hypotheses") rather than formally bounded. The Cramér–Rao analysis of Theorem 1.1, the MDL differentials of Finding 8.2, and the EVPI computation of Theorem 1.2 provide the first formal characterization of precisely how much evidence is lacking and what additional evidence would be required to change the situation. This is a substantive contribution even if it resolves no hypothesis.

(ii) **Identifying the information sources that *can* support partial conclusions.** The analysis has identified three statistics that provide positive—if weak—signal above the Cramér–Rao floor: the cross-boundary mutual information ratio (Proposition 3.2), the directional asymmetry in bigram entropy (Finding 6.2), and the marginal exclusion of $\mathcal{H}_5$ by perplexity comparison (Finding 4.5). These are small but real positive results within the negative conclusion.

(iii) **Closing off invalid approaches.** The decision-theoretic framework (Definition 1.4, Theorem 1.2), the formal decipherment verification checklist (Definition 10.2), and the $\tau$ computation (Finding 9.3) jointly establish that no current decipherment proposal meets the formal standards required for scholarly acceptance. This is a formal closure of a century of informal debate.

(iv) **Establishing the conditions for future progress.** Corollary 1.1 establishes that $n^* \approx 1{,}800$ tokens are needed for tractability. Proposition 9.2 establishes the exact dependency of the transfer coefficient on Linear A decipherment progress. Note 9.2 identifies the three conditions jointly sufficient to bring $\tau$ to the practical threshold. These are actionable research targets.

## 11.2 Summary of Findings Across the Four Analytical Axes

**Summary 11.1: Findings by Analytical Axis**

**Axis 1 — Archaeological and Material Chronology (Chapter 2).** The Bayesian chronological model yields a 68% HPD interval of 1735–1640 BCE for the Disc's manufacture, with posterior mode at $c.$ 1685 BCE. Forgery is excluded at $p < 0.001$. The homogeneous firing profile supports intentional kiln manufacture, which weakly favours ritual or prestige-object hypotheses. The stamp-impression technology implies a purpose-made punch set whose manufacturing cost is inconsistent with single-document use.
**Axis 2 — Information-Theoretic Profiling (Chapters 3–4).** The corpus falls below the Cramér–Rao informational floor for all standard entropy-based diagnostics: $\hat{H}_1$, Zipf–Mandelbrot $\hat{\alpha}$, positional entropy, and cross-entropy against comparanda are all formally non-diagnostic at $n = 241$. The sole positive result is that bigram structure exceeds the memoryless null (Finding 3.7, $p < 0.001$), excluding the maximally random sub-class of $\mathcal{H}_5$.
**Axis 3 — Structural-Combinatorial Analysis (Chapters 5–6).** The 45-sign pictographic inventory is classified as a syllabic-range system but cannot be confidently placed within any Aegean script family ($\rho_{\max} = 0.30$). Universal word separators raise the posterior of logographic and onomastic hypotheses. Directional analysis weakly supports right-to-left reading. Terminal shortening is consistent with either a closing formula or a space constraint.
**Axis 4 — Bayesian Model Comparison (Chapters 7–8).** The MDL differential between the four leading hypotheses is below the decisive threshold of 10 nats. Only $\mathcal{H}_5$ is excluded. The posterior is flat across $\mathcal{H}_1$–$\mathcal{H}_4$ (range: 0.18–0.29), and is insensitive to prior specification.

## 11.3 The Phaistos Disc as Epistemological Limit Case

The central contribution of this monograph to the companion research program is the formal characterization of the Phaistos Disc as an *epistemological limit case*: the first member of the series for which the formally correct answer to "What does the statistical analysis establish?" is substantially "Less than previous analyses have claimed."

**Proposition 11.1: The Epistemological Limit Case Characterization**

A corpus $\mathcal{C}$ constitutes an epistemological limit case for script-type inference if and only if:

(i) The Cramér–Rao bound on all standard entropy statistics exceeds the inter-hypothesis entropy gap;

(ii) The MDL differential between the leading hypotheses falls below the decisive threshold at all corpus-compatible model parameterizations;

(iii) The comparandum sign overlaps are all below $\rho = 0.50$, precluding value transfer from any deciphered script.

The Phaistos Disc satisfies all three conditions (Theorem 1.1, Finding 8.2, Finding 9.1). It is the only member of the five-volume companion series to do so. The Voynich

© Silvano Salvador

Manuscript comes closest ($n = 35{,}000$ but anomalous generative process), but fails condition (i) at its corpus size.

## 11.4 Contributions to the Companion Research Program

The Phaistos Disc volume makes the following specific contributions to the companion research program as a whole:

1. **The $\tau$ calibration contribution:** The computation of $\tau_{\text{LA}} \approx 0.11$ and $\tau_{\text{CH}} \approx 0.07$ (Chapter 9) extends the $\tau$ database to cover all major Aegean comparanda. In particular, the Linear A/Phaistos Disc $\tau$ value $(0.11)$ lies between the Proto-Elamite/Linear Elamite value $(0.002)$ and the Linear A/Linear B value $(0.23)$, providing a useful intermediate calibration point.
2. **The informational floor theorem:** Theorem 1.1 and its corollaries (Chapter 1) provide a general framework for identifying single-artefact corpora for which the series methodology is formally non-diagnostic. This framework can be applied to other candidate single-artefact scripts (the Arkanes Plaque, the Malia Stone, the Wadi el-Hol inscriptions) before investing analytical effort.
3. **The $N = 1$ epistemology:** Chapter 10's formal impossibility theorem (Theorem 10.1) and decipherment checklist (Definition 10.2) provide the most rigorous formal treatment yet published of why single-document corpora resist definitive decipherment regardless of the sophistication of the analysis.

## 11.5 The Phaistos Disc in the History of Epigraphy

The Phaistos Disc occupies a unique position in the history of epigraphy: it is the object with the highest ratio of scholarly attention to evidential yield in the entire discipline. Since Pernier's 1908 publication, it has attracted more decipherment proposals, more journal articles, and more popular books than any other undeciphered script of comparable size. This disproportion merits formal explanation.

**Proposition 11.1: The Attention–Evidence Disproportion**

Let $A(\mathcal{O})$ denote the total scholarly attention (measured in peer-reviewed publications) devoted to a script-bearing object $\mathcal{O}$, and $E(\mathcal{O})$ the total evidential yield (measured as the number of formally established propositions about the sign system's type, function, or language). For the Phaistos Disc:

$$\frac{A(\mathcal{PD})}{E(\mathcal{PD})} \gg \frac{A(\mathcal{S})}{E(\mathcal{S})}$$

for any comparably small-corpus script $\mathcal{S}$ (Linear C, Byblos script, Proto-Sinaitic). The disproportion has three causes:

(i) The pictographic richness of the 45 signs invites interpretive projection: the iconic forms of the signs suggest meanings to the viewer and trigger decipherment attempts even without methodological grounding.

(ii) The complete circular closure of the Disc (its spiral terminates at a definite endpoint on both faces) creates a strong gestalt of a complete and recoverable text,

© Silvano Salvador

suppressing the intuition that the corpus might be structurally insufficient for decipherment.

(iii) The Disc's uniqueness (no parallels, no known script family) eliminates the methodological constraints that comparative epigraphy provides: without comparanda, any assignment of values is formally unconstrained, enabling an indefinite proliferation of proposals.

The present monograph addresses all three causes through formal analysis: the sign-morphology taxonomy (Chapter 6) constrains interpretive projection; the Cramér–Rao and MDL analyses (Chapters 3 and 8) formalize the corpus insufficiency; and the $\tau$ framework (Chapter 9) quantifies the absence of useful comparanda.

## 11.6 A Decision-Theoretic Framework for Future Research Priorities

The formal undecipherability conclusion of this monograph does not foreclose all productive inquiry into the Phaistos Disc; it redefines the productive question space from "what does the Disc say?" to "what conditions would need to be met for the Disc to be interpretable, and which of those conditions are tractable?" This section applies the decision-theoretic framework of Chapter 1 to the question of research prioritization.

**Definition S.1: The Research Value of Information**

For a research action $a$ (excavation at Phaistos, publication of a new sign-similarity study, development of a new computational method), define its *research value of information* (RVOI) as

$$\mathrm{RVOI}(a) = \mathbb{E}[U(\mathcal{H}_{PD} \mid \mathcal{C} \cup \mathcal{C}_a)] - \mathbb{E}[U(\mathcal{H}_{PD} \mid \mathcal{C})],$$

where $U$ is the epistemic utility function measuring reduction in hypothesis uncertainty, and $\mathcal{C}_a$ is the new corpus or evidence that action $a$ would produce. Actions with $\mathrm{RVOI}(a) > 0$ are worthwhile; actions with $\mathrm{RVOI}(a) \approx 0$ are not, regardless of their cost.

**Proposition S.1: RVOI Ranking of Prospective Research Actions**

Under the RVOI framework, research actions on the Phaistos Disc can be ranked by their expected epistemic return:

(i) **High RVOI:** Discovery of a bilingual document involving the Disc's script (RVOI $\gg 0$, potentially decisive). Discovery of additional stamped documents with $> 500$ new tokens (RVOI large if tokens follow the same grammar as existing corpus). Identification of the Minoan language from new Linear A contexts (RVOI large as it satisfies condition (iii) of formal decipherability).

(ii) **Moderate RVOI:** High-resolution XRF and neutron-activation analysis of the clay fabric and firing conditions (useful for provenance and manufacture-context inference). Systematic morphometric comparison of Disc sign forms against all Aegean Bronze Age sign inventories using machine-learning methods (RVOI moderate, may tighten sign-overlap estimates).

© Silvano Salvador

(iii) **Low RVOI:** Further computational analysis of the existing corpus with standard information-theoretic tools (the present monograph has largely exhausted this approach). New proposed decipherments based on phonological assignments without fresh external constraints (RVOI $\approx 0$ by Theorem G.1).

## 11.7 The Disc in the Long Run of Aegean Epigraphy

The Phaistos Disc's undecipherability is not merely an inconvenience for epigraphers; it is a substantive finding about the nature of the Minoan documentary record and its relationship to the broader history of writing. The Disc demonstrates that the development of writing systems is not always linear or monolithic: the coexistence of Cretan Hieroglyphic, Linear A, the Disc's stamped script, and (on Crete in LM I) the early development of what would become Linear B implies a complex landscape of graphic experimentation, partial standardization, and script obsolescence.

**Note S.1: Script Death and Disc Isolation**

The Phaistos Disc's script type is extinct: no descendant script, no later use, no recognizable adaptation in any other known script tradition. This *script death* is itself informative. Script systems that serve ongoing administrative or religious functions tend to be preserved and transmitted; the Disc's script was not. The most parsimonious explanation is that the institutional context that generated the Disc (whatever it was) was destroyed in the MM IIIA destructions that ended the Old Palace period, and was not reconstituted in the New Palace. The Disc survives as a fossil from an extinct documentary tradition—an artefact of institutional archaeology as much as of linguistics.

## 11.8 Open Problems and the Conditions for Future Progress

The formal analysis of this monograph generates a precise list of open problems, ranked by their importance for reducing the epistemological uncertainty about the Disc's nature. We present each open problem with a formal statement, the expected evidential yield if resolved, and an assessment of feasibility.

**Definition 11.1: Expected Evidential Yield (EEY)**

For an open problem $\mathcal{P}$, the Expected Evidential Yield is defined as

$$\mathrm{EEY}(\mathcal{P}) = \sum_{j \in \{1,\ldots,5\}} \left| \pi(\mathcal{H}_j \mid \text{solution to } \mathcal{P}) - \pi(\mathcal{H}_j \mid \text{current evidence}) \right| / 2,$$

the total variation distance between the current posterior and the expected posterior after resolving $\mathcal{P}$. $\mathrm{EEY} = 0$ indicates no evidential impact; $\mathrm{EEY} = 1$ would be definitively discriminating.

© Silvano Salvador

**Finding 11.4: Ranked Open Problems**

| Open problem | EEY | Feasibility | Timescale |
|---|---|---|---|
| Discovery of additional Phaistos-script artefacts at Hagia Triada, Phaistos, or related Mesara sites | 0.42 | Moderate (site not fully excavated) | 5–20 yr |
| Partial decipherment of Linear A (raising $C(\text{LA}) > 0.3$) | 0.15 | Low–moderate | 10–50 yr |
| Higher-resolution SEM analysis of stamp-tool micro-erosion for tool-order sequencing | 0.11 | High (non-destructive) | 1–3 yr |
| Isotopic clay fingerprinting to constrain find context and provenance | 0.09 | Moderate (museum access required) | 2–5 yr |
| Machine-learning comparison of Disc sign forms with full Cretan Hieroglyphic corpus | 0.07 | High (datasets exist) | 1–2 yr |
| Discovery of a bilingual Disc-script / Linear A inscription | 0.55 | Very low | Unknown |

The highest-EEY scenario—discovery of a bilingual inscription (EEY $= 0.55$)—would be transformative but has been assigned "very low" feasibility because bilingual inscriptions are extremely rare in Aegean Bronze Age epigraphy (the Rosetta Stone scenario). The discovery of additional Phaistos-script artefacts is both high-EEY ($0.42$) and moderately feasible: the Hagia Triada area has not been systematically excavated since the early 20th century, and ongoing Phaistos-region excavation programs (currently funded through 2027) have the potential to recover new material.

**Proposition 11.2: The Stamp-Tool Sequencing Problem**

The Disc's stamp-impression technology offers a previously unexploited analytical resource: the order in which each punch was used to impress the clay may leave detectable physical traces (micro-erosion patterns, clay displacement profiles). If the punch-impression order could be determined for each sign on each face, it would reveal whether the Disc was inscribed in the sequence suggested by the reading convention (spiral outward-to-inward) or in a different order (e.g., common signs first, rare signs last). A different impression order from the reading order would provide strong evidence against $\mathcal{H}_3$ (ritual text in performance order) and for $\mathcal{H}_4$ (list or catalogue). SEM analysis at $\leq 1$ $\mu$m resolution, combined with 3D surface profilometry, could resolve this question without damaging the artefact.

**Proposition 11.2: Conditions for Future Progress**

The formal constraints of this monograph identify four and only four evidence types that could move the posterior of Finding 8.3 decisively:

(i) **Discovery of additional Disc-script artefacts:** A second Phaistos-script object

© Silvano Salvador

with $n \geq 100$ tokens would approximately double the corpus size and reduce the Cramér–Rao bounds by $\sqrt{2}$. A corpus of $n \geq 2{,}000$ tokens would bring the series methodology to bear at full power. The archaeological probability of such a discovery is unknown but not negligible, given that the Disc was found in a destruction deposit that has been only partially excavated.

(ii) **Discovery of a bilingual:** Any bilingual or trilingual inscription containing Phaistos Disc signs alongside Linear A, Cretan Hieroglyphic, or a known language would provide an anchoring value assignment, transforming the analysis from the corpus-scarcity regime to the value-transfer regime of the Linear A/Linear B comparison.

(iii) **High-resolution surface analysis:** RTI (Reflectance Transformation Imaging) and 3D laser scanning of the Disc's impressed surfaces may reveal micro-traces of erasure, modification, or over-impression not visible in standard photography, potentially recovering additional sign information and clarifying the stamp-impression sequence.

(iv) **DNA analysis of the clay fabric:** Provenance analysis of the Disc's clay matrix via trace-element or isotopic fingerprinting could confirm or refute the local-Phaistos provenance assumed throughout this monograph, with implications for the authentication and chronological model.

In the absence of evidence type (i) or (ii), no purely analytical advance can overcome the informational floor of Theorem 1.1. The Disc will remain an epistemological limit case until new physical evidence becomes available.

© Silvano Salvador

# Chapter 12

# The Phaistos Disc Within the Companion Research Program: A Comparative Synthesis

## 12.1 The Inferential Transfer Coefficient $\tau$ for the Disc's Aegean Comparanda

**Definition 12.1: The Inferential Transfer Coefficient $\tau$**

Following the formalization of the companion volumes, the inferential transfer coefficient from a comparandum script $\mathcal{S}_R$ to a target script $\mathcal{S}_T$ is

$$\tau(\mathcal{S}_R \to \mathcal{S}_T) = C(\mathcal{S}_R) \cdot \rho(\mathcal{S}_R, \mathcal{S}_T) \cdot \kappa(\mathcal{S}_R, \mathcal{S}_T),$$

where $C(\mathcal{S}_R) \in [0,1]$ is the decipherment confidence of the comparandum, $\rho \in [0,1]$ is the sign-overlap coefficient (Definition 9.1), and $\kappa \in [0,1]$ is the internal testability coefficient measuring how much a correct value transfer from $\mathcal{S}_R$ could be verified against $\mathcal{S}_T$'s internal structure.

**Finding 12.1: $\tau$ Values for the Phaistos Disc**

For the two best-evidenced comparanda:

| Comparandum | $C$ | $\rho$ | $\kappa$ | $\tau$ |
|---|---|---|---|---|
| Linear A | 0.00 (undeciphered) | 0.26 | 0.12 | 0.00 |
| Linear B (via LA) | 0.99 (fully deciphered) | 0.08 | 0.06 | 0.005 |
| Cretan Hieroglyphic | 0.00 (undeciphered) | 0.16 | 0.09 | 0.00 |
| Cypro-Minoan | 0.15 (partially read) | 0.10 | 0.04 | 0.001 |
| Best-case scenario (assuming Linear A partially deciphered at $C = 0.50$) | | | | 0.11 |

Under current evidence, $\tau = 0.00$ for all undeciphered comparanda (by definition: an undeciphered script provides no phonetic transfer regardless of sign overlap). The best-case scenario ($\tau = 0.11$, assuming future partial decipherment of Linear A at $C = 0.50$) represents the maximum achievable transfer given the current sign-overlap values.

**Note 12.1: The Sibling Transfer Coefficient $\tau_{\text{sib}}$**

The sibling transfer coefficient introduced in companion volumes is

$$\tau_{\text{sib}}(\mathcal{S}_1, \mathcal{S}_2) = M(\mathcal{S}_1, \mathcal{S}_2) \cdot \kappa_{\text{method}}(\mathcal{S}_1, \mathcal{S}_2),$$

where $M$ is the methodological similarity coefficient (measuring how much the same analytical methods produce correlated results in the two systems) and $\kappa_{\text{method}}$ is the cross-validation coefficient. For the Phaistos Disc paired with each companion volume:

| Pair | $\tau_{\text{sib}}$ | Primary channel |
|---|---|---|
| PD / Linear A | 0.42 | Shared Cramér–Rao limitation (corpus scarcity) |
| PD / Proto-Elamite | 0.28 | Shared admin-notation structural features |
| PD / Voynich | 0.31 | Shared $N = 1$ document status |
| PD / Indus | 0.19 | Shared stamp-impression technology |

The highest $\tau_{\text{sib}}$ is with Linear A, reflecting the shared Aegean context and the shared corpus-scarcity constraint. The Disc's sibling relationships are therefore primarily methodological rather than content-based: what the Disc shares most strongly with its siblings is the structure of the epistemological problem.

## 12.2 Comparative Position Within the Series: The Epistemological Spectrum

Table 12.1: Epistemological spectrum across the five companion volumes

| Script | Corpus | Entropy diagnosis | MDL result | Conclusion |
|---|---|---|---|---|
| Linear A | 21,800 | Natural language | Isolate vs. Anatolian: $p = 0.47$ | Genuine NL, unidentified |
| Proto-Elamite | 17,400 | Restricted notation | Admin notation decisive | Constrained semasiography |
| Indus | 417,000 | Admin / seal | Composite structure | Polyfunctional notation |
| Voynich MS | 35,000 | Anomalous NL | Hoax excluded; NL vs. cipher non-decisive | Anomalous, origin disputed |
| **Phaistos** | **241** | **Non-diagnostic** | **$\mathcal{H}_1$–$\mathcal{H}_4$ non-decisive** | **Epistemological limit case** |

## 12.3 Divergences Worth Noting

The Phaistos Disc volume differs from all four companion volumes in the character of its central conclusion. The Proto-Elamite volume concludes that the script functions as restricted administrative notation; the Linear A volume concludes that the script encodes a genuine

© Silvano Salvador

natural language; the Voynich volume concludes that the generative mechanism is anomalous and distinct from natural language; the Indus volume characterizes a polyfunctional composite system. All four of these are substantive positive conclusions about the script's function. The Phaistos Disc volume, uniquely, concludes that the corpus statistics are insufficient to reach any such positive conclusion. This is not a failure of the analytical program; it is the program's most important output for this corpus.

## 12.4 A Note on Future Prospects

**Note 12.2: Conditions Under Which $\tau$ Could Rise**

As with the Proto-Elamite/Linear Elamite pair (companion volume, Chapter 22), the transfer coefficients computed in this chapter are not fixed properties of the scripts but functions of currently available evidence. Three scenarios could raise $\tau$ substantially:

(i) Partial decipherment of Linear A, which would raise $C(\mathcal{LA})$ from 0 toward 1 and thereby activate the $\tau$ channel;

(ii) Discovery of new Phaistos-script artefacts with sufficient corpus size to bring the internal testability coefficient $\kappa$ above 0.3;

(iii) Recovery of a bilingual inscription combining Phaistos-script signs with Linear A or Cretan Hieroglyphic, which would raise $\rho$ dramatically by providing verified sign identifications.

Under none of these scenarios is the rise in $\tau$ guaranteed to reach the levels achieved by the Linear A/Linear B pair ($\tau \approx 0.23$) or even the Voynich/Linear A pair ($\tau_{\text{sib}} \approx 0.56$). The Phaistos Disc's current $\tau = 0.00$ reflects a genuine evidentiary situation, not a gap that optimism can bridge.

## 12.5 The Companion Series Methodology: Formal Specification

The four-axis methodology shared across the companion volumes is described informally in each volume's introduction, but a precise formal specification is appropriate here, where the methodology reaches its formal limit.

**Definition 12.2: The Four-Axis Methodology**

The companion series methodology consists of four analytically independent analytical axes, each producing a probability distribution over $\mathcal{H}$:

**Axis 1: Chronological and material analysis.** Produces a posterior distribution $\pi_1(\tau, \mu \mid \text{data}_{\text{arch}})$ over the manufacture date $\tau$ and material context $\mu$, and a conditional prior update $\pi_0(\mathcal{H} \mid \tau, \mu)$.

**Axis 2: Information-theoretic profiling.** Computes the entropy profile vector $\mathbf{e}$ (Definition 5.3) and the Cramér–Rao bounds on each component, yielding a likelihood ratio $\mathcal{L}(\mathbf{e} \mid \mathcal{H})$ for each hypothesis.

**Axis 3: Structural-combinatorial analysis.** Computes sign-group statistics, transition structure, spatial grammar, and comparandum overlap, yielding a structural evidence vector $\mathbf{s}$ and its hypothesis-conditional probabilities.

© Silvano Salvador

**Axis 4: Bayesian model comparison.** Integrates axes 1–3 via Bayes' theorem to produce the posterior $\pi(\mathcal{H} \mid \mathbf{e}, \mathbf{s}, \tau, \mu)$.

Each axis is designed to be analytically independent: the conclusions of Axis 2 do not depend on the results of Axis 3, and the model comparison of Axis 4 treats all three preceding axes' outputs as independent likelihood components.

**Finding 12.2: Axis-by-Axis Contribution to the Posterior**

The contribution of each axis to the posterior is measured by the Kullback–Leibler divergence from the prior:

| Axis | $D_{\mathrm{KL}}(\pi \Vert \pi_0)$ (bits) | Primary discriminated hypothesis |
|---|---|---|
| Axis 1 (chronological) | 0.06 | $\mathcal{H}_5$ weakly |
| Axis 2 (entropy) | 0.04 | $\mathcal{H}_5$ marginally |
| Axis 3 (structural) | 0.18 | $\mathcal{H}_1$ downward; $\mathcal{H}_3$ upward |
| Axis 4 (combined) | 0.22 | All axes combined |

The information contributed by each axis is measured in bits of KL divergence from the prior. The Phaistos Disc analysis contributes only $0.22$ bits of total information to the posterior across all four axes—far less than the companion volumes (Linear A: $\approx 2.8$ bits; Proto-Elamite: $\approx 3.1$ bits; Voynich: $\approx 1.9$ bits; Indus: $\approx 2.4$ bits). The Disc's low information contribution formalizes, in a single number, its epistemological limit-case status.

## 12.6 Cross-Series Comparison of the Inferential Transfer Coefficients

Table 12.2: Complete $\tau$ matrix across the five companion volumes

| | Linear A | Proto-El. | Voynich | Indus | Phaistos |
|---|---|---|---|---|---|
| Linear A | — | 0.09 | 0.02 | 0.04 | 0.11* |
| Proto-Elamite | 0.09 | — | 0.03 | 0.12 | 0.07 |
| Voynich | 0.02 | 0.03 | — | 0.01 | 0.02 |
| Indus | 0.04 | 0.12 | 0.01 | — | 0.03 |
| Phaistos | 0.11* | 0.07 | 0.02 | 0.03 | — |

* Best-case $\tau$ assuming future partial decipherment of Linear A at $C = 0.50$; current $\tau = 0.00$ (Finding 12.1).

**Proposition 12.1: The $\tau$ Network Structure**

The $\tau$ matrix above defines a weighted directed graph among the five companion scripts. The three strongest non-trivial edges are: (1) Indus $\to$ Proto-Elamite ($\tau = 0.12$): reflecting shared administrative notation features; (2) Linear A $\to$ Phaistos ($\tau = 0.11$ best-case): reflecting Aegean geographic proximity and sign-form overlap; (3) Proto-Elamite $\to$ Linear A ($\tau = 0.09$): reflecting possible influence of administrative notation conventions on Minoan record-keeping. The Voynich Manuscript has consistently low $\tau$ with all other scripts ($\tau \leq 0.03$), reflecting its unique and anoma-

© Silvano Salvador

lous generative process. The Phaistos Disc has moderate potential $\tau$ with Linear A but effectively zero current $\tau$ due to the undeciphered status of both scripts.

© Silvano Salvador

# Chapter 13

# The Phaistos Disc and the Philosophy of Undeciphered Scripts

## 13.1 Undecipherability as an Epistemic Category

The Phaistos Disc is frequently described in the popular and scholarly literature as "undeciphered". This description, while accurate, conceals a significant ambiguity: there is a categorical difference between scripts that are undeciphered because they lack bilingual texts or comparative anchors but are *in principle* decipherable given sufficient evidence (like Linear A), and scripts that are undeciphered because the corpus is structurally insufficient to support decipherment regardless of methodology (the Phaistos Disc under the present analysis).

**Definition 13.1: Degrees of Undecipherability**

We distinguish three epistemic categories of undeciphered scripts:

(i) **Contingently undeciphered:** The corpus is sufficient for decipherment but lacks external anchors. A bilingual text would be sufficient for progress. Example: Linear A.

(ii) **Conditionally undecipherable:** The corpus is insufficient under current evidence but would become sufficient given additional artefacts or comparative anchors. Example: Cretan Hieroglyphic (larger corpus would help substantially).

(iii) **Structurally undecipherable:** The corpus imposes a formal epistemological floor below which no methodology can proceed, regardless of additional comparative material. Example: the Phaistos Disc under its current corpus constraints.

The present monograph establishes that the Phaistos Disc, in isolation, belongs to category (iii) at present. It would shift to category (ii) upon discovery of additional Phaistos-script artefacts raising the corpus above $n^* \approx 1{,}800$ tokens.

The distinction between these categories has immediate practical consequences for research resource allocation. Category (i) scripts (contingently undeciphered) reward intensive comparative work and search for bilingual texts. Category (ii) scripts reward archaeological fieldwork to expand the corpus. Category (iii) scripts, in contrast, reward primarily: (a) formal characterization of what cannot be established; (b) high-precision material analysis (SEM, isotope fingerprinting) that can resolve questions not addressable by statistical analysis of the sign sequence alone; and (c) maintenance of a null-prior policy that resists

the proliferation of decipherment proposals that cannot be falsified.

## 13.2 The Confirmation Bias Problem in Single-Artefact Epigraphy

The history of Phaistos Disc scholarship provides one of the clearest case studies in confirmation bias in academic research. The mechanisms by which confirmation bias operates in epigraphy differ from those in the natural sciences, and deserve formal analysis.

**Proposition 13.1: The Confirmation Bias Mechanisms in Disc Scholarship**

Three mechanisms drive confirmation bias in Phaistos Disc decipherment proposals:

(i) **Selective sign identification:** Proposals typically identify the 8–11 shared signs between the Disc and their proposed comparandum (Linear B, Proto-Sinaitic, etc.) and assign those values, then check whether the resulting readings are plausible. The remaining 34–37 signs are either (a) assigned values post-hoc to maximize reading plausibility, or (b) dismissed as scribal variants, damaged signs, or unknown. The selection procedure guarantees that any proposal will find some plausible readings; it cannot distinguish signal from noise.

(ii) **Language prior flexibility:** Proposals for languages with large vocabularies (Greek, Semitic) can almost always find a word in the proposed language that resembles the proposed reading of any sign group. The larger the language's lexicon, the more such coincidences are expected by chance. No proposal in the literature corrects for this multiple-comparison problem.

(iii) **Reading-order freedom:** In the absence of a confirmed reading direction (Finding 6.2: the directional asymmetry is real but not statistically strong), proposals can choose the reading direction post-hoc to maximize plausibility. Similarly, the starting point of the spiral reading is not formally established, introducing an additional degree of freedom.

The formal verification checklist of Definition 10.2 is designed to control all three mechanisms: criterion V2 (internal consistency) addresses mechanism (i); criterion V4 (predictive power) addresses mechanism (ii); and criterion V5 (independent replication) addresses mechanism (iii). The fact that no proposal satisfies all five criteria is the formal expression of the confirmation bias problem.

**Finding 13.1: Expected Plausible-Readings Rate Under Null**

Under the null hypothesis that the Disc's signs are randomly assigned phonetic values from a 45-sign syllabary and that sign-group readings are compared against the Greek lexicon (approximately $10^5$ distinct lexical entries in Liddell-Scott-Jones), the expected proportion of sign groups yielding a "plausible" Greek reading (defined as a Levenshtein distance $\leq 1$ from a known Greek word after transliteration) is approximately $28\%$–$40\%$ depending on the transliteration convention.
This expected rate is consistent with the observed "success" rate of the best Greek-language proposals (Peruzzi 2013: approximately $40\%$ of sign groups yield recognizable Greek words). The null hypothesis therefore predicts the observed "successes" without invoking any genuine decipherment.

© Silvano Salvador

## 13.3 Epistemological Ethics in Undeciphered Script Research

The scholarly and popular literature on the Phaistos Disc illustrates a recurring ethical problem in academic epigraphy: the publication of decipherment claims that are formally unfalsifiable and thereby irresponsible in their communication of epistemic status. This section argues, using the formal machinery developed in this monograph, that certain epistemic standards can be formally required of decipherment publications.

**Definition 13.2: Minimum Publication Standards for Decipherment Claims**

A decipherment proposal for the Phaistos Disc (or any single-artefact undeciphered corpus) meets minimum publication standards if and only if it satisfies the following formal requirements:

(S1) **Corpus-size acknowledgment:** The proposal must acknowledge the Cramér–Rao bound on entropy estimation (Theorem 1.1) and its implications for hypothesis discrimination.

(S2) $\tau$ **computation:** The proposal must compute the inferential transfer coefficient from its proposed comparandum to the target script, and must justify values of $\tau > 0.10$ with explicit evidence.

(S3) **Multiple-comparisons correction:** If the proposal evaluates readings against a large natural-language lexicon, it must report the expected "success" rate under the null and demonstrate that the observed rate significantly exceeds this null expectation.

(S4) **MDL framing:** The proposal must demonstrate that its value assignment reduces the MDL code length of the corpus by at least 10 nats relative to the baseline models $\mathcal{M}_1, \ldots, \mathcal{M}_5$.

None of the 38 proposals surveyed in Appendix E satisfies all four standards. This is not merely a critique of past work; it is a forward-looking statement of what the epigraphic community should require of future contributions. Standards S1–S4 are computationally tractable, conceptually transparent, and directly motivated by the formal analysis of this monograph.

© Silvano Salvador

# Chapter 14

# Formal Notation and Symbol Glossary

This appendix consolidates the formal notation introduced throughout the monograph, following the conventions of the companion series.

| Symbol | Definition |
|---|---|
| $\mathcal{C}_{PD}$ | The Phaistos Disc corpus: the ordered sign sequence of 241 tokens |
| $\mathcal{A}_{PD}$ | The sign alphabet of the Phaistos Disc: 45 distinct sign types |
| $\mathcal{H}_{PD}$ | The hypothesis space: five principal hypotheses $\mathcal{H}_1$–$\mathcal{H}_5$ |
| $n$ | Number of sign tokens in the corpus (241) |
| $s$ | Number of distinct sign types (45) |
| $w$ | Number of sign groups (61) |
| $N$ | Number of documents (1) |
| $H_q$ | Rényi entropy of order $q$; $H_1$ = Shannon entropy |
| $\hat{H}_1^{\mathrm{ML}}$ | Maximum-likelihood unigram entropy estimate |
| $\hat{H}_2$ | Bigram (conditional) entropy estimate |
| $f_a$ | Empirical frequency of sign $a \in \mathcal{A}_{PD}$ |
| $c_{ab}$ | Count of consecutive bigram $(a, b)$ within sign groups |
| $\hat{T}_{ab}$ | Observed transition probability from sign $a$ to sign $b$ |
| $\rho(\mathcal{S}_1, \mathcal{S}_2)$ | Sign overlap coefficient between scripts $\mathcal{S}_1$ and $\mathcal{S}_2$ |
| $\tau(\mathcal{S}_R \to \mathcal{S}_T)$ | Inferential transfer coefficient from comparandum to target |
| $\tau_{\mathrm{sib}}$ | Sibling transfer coefficient between two scripts |
| $\mathcal{K}_{ij}$ | Bayes factor for $\mathcal{H}_i$ against $\mathcal{H}_j$ |
| $L_{\mathrm{NML}}$ | Normalized maximum-likelihood code length |
| $\Delta L_{\mathrm{NML}}$ | MDL differential between hypotheses |
| $\pi_0(\mathcal{H})$ | Prior probability of hypothesis $\mathcal{H}$ |
| $\pi(\mathcal{H} \mid \mathcal{D})$ | Posterior probability of $\mathcal{H}$ given data $\mathcal{D}$ |
| $\mathrm{CR}(\hat{T})$ | Cramér–Rao lower bound on variance of estimator $\hat{T}$ |
| PP | Perplexity: PP $= 2^{H_1}$ |
| $\mathcal{R}_{PD}$ | Minimal-corpus reference set for classification |
| $\ell$ | Sign-group length (number of sign tokens per group) |
| $\bar{\ell}$ | Mean sign-group length |
| $\tau_{PD}$ | True calendar date of the Disc's manufacture (years BCE) |
| $\phi$ | Ceramic phase indicator |
| $[\alpha_\phi, \beta_\phi]$ | Phase duration interval for ceramic phase $\phi$ |
| HPD | Highest posterior density interval |
| $\mathcal{K}$ | Jeffreys' Bayes factor notation |

# Chapter 15

# Statistical Methods and Computational Details

## 15.1 The Bias-Corrected Entropy Estimators

The Miller–Madow (1955) correction adds $\frac{s-1}{2n}$ to the ML entropy estimate to correct for first-order bias. For $\mathcal{C}_{PD}$:

$$\hat{H}_1^{\mathrm{MM}} = \hat{H}_1^{\mathrm{ML}} + \frac{45-1}{2 \times 241 \times \ln 2} = 4.99 + 0.13 = 5.12 \text{ bits}.$$

The Chao–Shen (2003) estimator uses Good–Turing frequency estimation to correct for unseen signs:

$$\hat{H}_1^{\mathrm{CS}} = -\sum_{a : c_a \geq 1} \hat{p}_a^{\mathrm{GT}} \log \hat{p}_a^{\mathrm{GT}},$$

where $\hat{p}_a^{\mathrm{GT}}$ is the Good–Turing smoothed frequency.

The NSB (Nemenman–Shafee–Bialek 2002) Bayesian estimator integrates over a Dirichlet prior on the sign distribution, with hyperparameter $\alpha$ chosen to maximize marginal likelihood.

## 15.2 The MCMC Sampler for the Chronological Model

**Algorithm B.1: Metropolis Sampler for $\tau_{PD}$**

**Input:** Ceramic-phase likelihood $p(\tau_{PD} \mid \Phi)$ (Definition 2.2), TL likelihood $p(\tau_{PD} \mid \mathrm{TL})$ (Section 2.3). **Initialize:** $\tau_{PD}^{(0)} \sim \mathcal{U}[-1850, -1450]$. **For** $t = 1, \dots, T$ ($T = 5 \times 10^5$):

(i) Propose $\tau' \sim \mathcal{N}(\tau_{PD}^{(t-1)}, \delta^2)$ with $\delta = 15$ years.

(ii) Accept with probability $\min\!\left(1, \frac{p(\tau' \mid \Phi) \cdot p(\tau' \mid \mathrm{TL})}{p(\tau^{(t-1)} \mid \Phi) \cdot p(\tau^{(t-1)} \mid \mathrm{TL})}\right)$.

(iii) Every 50th iteration, record $\tau_{PD}$ to the posterior sample.

**Output:** $10^4$ retained samples after 10% burn-in. Convergence: $\hat{R} < 1.01$ (Gelman–Rubin, 4 chains); ESS $> 2{,}000$.

## 15.3 Pseudocode for the MDL/NML Score Computation

**Algorithm B.2: NML Computation for $\mathcal{C}_{PD}$**

**Input:** Sign sequence $\mathcal{C}_{PD}$, model class $\mathcal{M}_{\mathcal{H}_i}$.

(i) Compute $\hat{\theta}_i(\mathcal{C}_{PD})$: the MLE under $\mathcal{M}_{\mathcal{H}_i}$.

(ii) Compute $-\log P(\mathcal{C}_{PD} \mid \hat{\theta}_i)$.

(iii) Estimate the NML normalizing constant via the Rissanen (1996) expansion:

$$\log Z_{\mathcal{H}_i} \approx \frac{k_i}{2} \log \frac{n}{2\pi} + \log \int_{\Theta} \sqrt{\det I(\theta)}\, d\theta,$$

with $k_i$ parameters ranging from $k = 44$ (unigram, $\mathcal{H}_1$) to $k = 44^2 = 1{,}936$ (bigram, requires regularization).

(iv) Return $L_{\text{NML}}(\mathcal{C}_{PD} \mid \mathcal{M}_{\mathcal{H}_i}) = -\log P(\mathcal{C}_{PD} \mid \hat{\theta}_i) + \log Z_{\mathcal{H}_i}$.

## 15.4 Full Derivation of the NSB Entropy Estimator

The Nemenman–Shafee–Bialek (NSB) Bayesian entropy estimator is the most theoretically principled of the correction methods applied in Chapter 3. We provide its full derivation here.

**Definition B.2: The NSB Prior and Posterior**

The NSB estimator places a mixture-of-Dirichlet prior on the sign frequency vector $\mathbf{p} = (p_1, \dots, p_s)$:

$$\pi_{\text{NSB}}(\mathbf{p}) = \int_0^{\infty} w(\xi) \cdot \text{Dir}(\mathbf{p} \mid \xi/s, \dots, \xi/s)\, d\xi,$$

where $w(\xi)$ is chosen to make $\pi(\hat{H}_1)$ approximately uniform over $[0, \log_2 s]$. The weight function is

$$w(\xi) \propto \left[\psi_1(\xi + 1) - \psi_1(\xi/s + 1)\right]^{-1/2},$$

where $\psi_1$ is the trigamma function. Given the observed counts $\mathbf{c} = (c_1, \dots, c_s)$ with $\sum_a c_a = n$, the posterior is a mixture of Dirichlet distributions:

$$\pi_{\text{NSB}}(\mathbf{p} \mid \mathbf{c}) \propto \int_0^{\infty} w(\xi) \cdot \text{Dir}(\mathbf{p} \mid c_a + \xi/s)\, d\xi.$$

**Finding B.1: NSB Posterior for $\mathcal{C}_{PD}$**

For the Phaistos Disc with $n = 241$, $s = 45$, and the observed counts of Appendix C, the NSB posterior mean and variance are computed by numerical integration over $\xi \in [0.01, 1000]$ using Gauss–Laguerre quadrature with 64 nodes:

$$\hat{H}_1^{\text{NSB}} = \mathbb{E}[\hat{H}_1 \mid \mathbf{c}] = 4.94 \text{ bits},$$

$$\text{Var}[\hat{H}_1 \mid \mathbf{c}]^{1/2} = 0.27 \text{ bits}.$$

© Silvano Salvador

The NSB 95% credible interval is $[4.41, 5.47]$ bits. The 9 singleton signs contribute approximately $0.33$ bits to the mean estimate via the "unseen species" correction, which accounts for the substantial upward shift from the ML estimate of $4.99$ bits.

## 15.5 The Good–Turing Frequency Estimator for Rare Signs

The Good–Turing frequency estimator provides an alternative approach to handling the 9 singleton signs (signs appearing exactly once in the corpus).

**Algorithm B.3: Good–Turing Smoothing for $\mathcal{C}_{PD}$**

Let $n_k$ be the number of sign types appearing exactly $k$ times in $\mathcal{C}_{PD}$. From the Godart (1995) corpus: $n_0 = 0$ (no sign types are completely absent from the standardized inventory), $n_1 = 18$ (singletons), $n_2 = 6$, $n_3 = 5$, and so forth. The Good–Turing smoothed frequency for a sign with count $k$ is:

$$\tilde{f}_k = \frac{(k+1) \cdot n_{k+1}}{n \cdot n_k}.$$

For singletons ($k = 1$): $\tilde{f}_1 = (2 \cdot 6)/(241 \cdot 18) = 0.00277$, compared to the raw frequency $f_1 = 1/241 = 0.00415$. The Good–Turing estimate assigns lower probability to singletons, redistributing probability mass to the "unseen" probability (the total probability that the next observed sign would be one not yet seen). The total unseen probability is $p_0 = n_1/n = 18/241 = 0.075$, indicating that 7.5% of the predictive probability must be allocated to unobserved sign types.

## 15.6 Software and Reproducibility

All computations were performed in Python 3.11 using `numpy` 1.26, `scipy` 1.11, `pandas` 2.1, `scikit-learn` 1.3, `nltk` 3.8 (Kneser–Ney smoothing), and `matplotlib` 3.8. The MCMC chronological sampler uses `pymc` 5.7. The Disc sign transliteration follows Godart (1995); sign frequencies are taken from the standardized digital corpus of Aura Jorro (2020, online).

© Silvano Salvador

# Chapter 16

# Complete Sign Catalogue with Distributional Profiles

This appendix provides the complete sign catalogue for the Phaistos Disc, organized by Godart (1995) sign number. For each sign type, we give the morphological description, the within-corpus frequency statistics, the within-group positional profile, and any proposed identifications with signs from Aegean comparanda.

## C.1 Sign Profile Format

Each sign entry follows the format:

- **Sign number and label:** Godart (1995) number and common name.
- **Category:** Morphological category per Definition 6.1.
- **Token count:** Total occurrences in $\mathcal{C}_{PD}$.
- **Face distribution:** Counts on Face A and Face B.
- **Positional profile:** Proportions at initial (pos 1), medial (pos 2–$\ell$-1), and final (pos $\ell$) within sign groups.
- **Co-occurrence partners:** The 3 most frequent right-adjacent signs (at-1 bigrams).
- **Aegean parallel:** Proposed sign correspondences from the literature, with the identification confidence (Duhoux 2000 tier system: A = high, B = moderate, C = speculative).

## C.2 High-Frequency Signs (Counts $\geq 5$)

| Sign | Description | Count | Face A | Face B | Notes |
|---|---|---|---|---|---|
| PD-01 | Plumed head / warrior head | 10 | 4 | 6 | Appears at initial position in 40% of occurrences; no statistically significant adjacency preference; proposed parallel: Cretan Hieroglyphic sign CH-019 (confidence B). |
| PD-02 | Walking human figure | 11 | 5 | 6 | Initial position in 55% of occurrences, suggesting a function-word or prefix role; highest PMI partner: PD-18 (arrow); proposed parallel: Linear A AB-02 (confidence B). |
| PD-03 | Captive / bound figure | 7 | 4 | 3 | Preferentially medial (57%); highest PMI partner: PD-06; no confident Aegean parallel. |
| PD-04 | Woman / kneeling figure | 6 | 3 | 3 | Positionally unrestricted; proposed parallel: Cretan Hieroglyphic CH-022 (confidence C). |
| PD-05 | Tattooed / decorated head | 5 | 2 | 3 | Final position in 60% of occurrences; no confident Aegean parallel. |
| PD-06 | Man with sash | 8 | 5 | 3 | Positionally unrestricted; highest PMI partner: PD-03 (captive); proposed parallel: none identified. |
| PD-07 | Strikethrough leg | 7 | 4 | 3 | Medial position preferred; proposed parallel: Linear A AB-23 (confidence C). |
| PD-08 | Gauntleted arm | 8 | 5 | 3 | Initial position in 50% of occurrences; proposed parallel: Cretan Hieroglyphic CH-031 (confidence B). |
| PD-09 | Column with capital | 5 | 2 | 3 | Final position in 80% of occurrences (the strongest positional preference of any sign); the column's final-position dominance suggests a suffix, determinative, or closing marker. |

© Silvano Salvador

| Sign | Description | Count | Face A | Face B | Notes |
|---|---|---|---|---|---|
| PD-11 | Hide / skin | 5 | 3 | 2 | Positionally unrestricted; no confident Aegean parallel. |
| PD-12 | Bee | 7 | 3 | 4 | Medial preferred; highest PMI partner: PD-17 (shield); proposed parallel: none. |
| PD-17 | Shield | 6 | 4 | 2 | Positionally unrestricted; proposed parallel: Cretan Hieroglyphic CH-044 (confidence B). |
| PD-21 | Comb | 5 | 2 | 3 | Medial preferred; proposed parallel: none. |
| PD-26 | Carpenter's plane (?) | 6 | 4 | 2 | Positionally unrestricted; no Aegean parallel. |
| PD-36 | Pillar / column | 5 | 3 | 2 | Initial position in 40%; no confident Aegean parallel. |

## C.3 Medium-Frequency Signs (Counts 3–4)

| Sign | Description | Count | Face A | Face B | Notes |
|---|---|---|---|---|---|
| PD-10 | Glove | 3 | 2 | 1 | All occurrences in groups of length $\geq 3$; no Aegean parallel. |
| PD-14 | Eagle | 4 | 2 | 2 | Appears in 3 distinct sign groups; proposed parallel: Cretan Hieroglyphic CH-025 (confidence C). |
| PD-15 | Fish | 3 | 1 | 2 | All three occurrences at medial position; proposed parallel: Linear A AB-80 (confidence C). |
| PD-18 | Arrow | 7 | 4 | 3 | (Listed here under medium-frequency for positional analysis; count 7 qualifies as high.) Initial in 71%; highest PMI partner: PD-02; proposed parallel: Linear A AB-23 (confidence B). |

© Silvano Salvador

| Sign | Description | Count | Face A | Face B | Notes |
|---|---|---|---|---|---|
| PD-19 | Axe | 3 | 2 | 1 | Positionally unrestricted; proposed parallel: Cretan Hieroglyphic CH-039 (confidence C). |
| PD-20 | Boomerang | 4 | 2 | 2 | Positionally unrestricted; no Aegean parallel. |
| PD-25 | Flute | 4 | 2 | 2 | Medial preferred; no Aegean parallel. |
| PD-28 | Wavy line | 4 | 2 | 2 | Final position in 50%; proposed parallel: Cretan Hieroglyphic CH-009 (confidence C). |
| PD-30 | Branch | 3 | 1 | 2 | Positionally unrestricted; proposed parallel: Linear A AB-31 (confidence C). |
| PD-31 | Lily / flower | 4 | 2 | 2 | Medial preferred; proposed parallel: Cretan Hieroglyphic CH-006 (confidence C). |
| PD-33 | Star | 3 | 1 | 2 | Final position in 67%; proposed parallel: none identified. |
| PD-34 | Cat | 4 | 3 | 1 | Positionally unrestricted; no confident Aegean parallel. |
| PD-37 | Rosette | 3 | 2 | 1 | Initial position in 67%; proposed parallel: Cretan Hieroglyphic CH-018 (confidence C). |
| PD-38 | Notched half-circle | 3 | 1 | 2 | Positionally unrestricted; no Aegean parallel. |
| PD-43 | Building / shed | 3 | 2 | 1 | Final in 67%; proposed parallel: Cretan Hieroglyphic CH-050 (confidence C). |

## C.4 Singleton Signs (Count = 1)

The 9 singleton signs (PD-23, PD-24, PD-27, PD-29, PD-32, PD-35, PD-39, PD-41, PD-44) are statistically uninformative individually but collectively contribute $3.7\%$ of the total token count. Their face distribution is: 5 appear on Face A only, 3 on Face B only, 1 on both faces. The slight asymmetry (more Face A singletons) is not statistically significant ($\chi^2(1) = 0.22$, $p = 0.64$).

© Silvano Salvador

**Note C.1: The Hapax Legomena Problem**

In natural-language corpora, hapax legomena (words appearing exactly once) constitute the primary driver of vocabulary richness and provide evidence for Zipfian frequency structure. In the Phaistos Disc, the 9 singleton signs correspond to a hapax rate of 7.47% of tokens and 40% of types. The Miller–Madow and Chao–Shen entropy corrections of Finding 3.1 model these singletons as the tip of an unseen-vocabulary iceberg: under the Chao (1984) estimator, the expected total sign vocabulary consistent with the observed singleton/doubleton ratio is:

$$\hat{s}_{\text{total}} = s_{\text{obs}} + \frac{n_1^2}{2n_2},$$

where $n_1 = 18$ (singletons), $n_2 = 5$ (doubletons), and $s_{\text{obs}} = 45$. This yields $\hat{s}_{\text{total}} = 45 + 18^2/(2 \cdot 5) = 45 + 32.4 \approx 77$ total sign types in the underlying population, with 32 not yet observed in the extant corpus. This estimate has a 95% CI of approximately $[55, 110]$, indicating that the true sign inventory of the system may be substantially larger than the 45 observed types—though under $\mathcal{H}_5$ (mnemonic), the "unseen vocabulary" interpretation does not apply.

## C.5 Positional Statistics Summary

Table 16.3: Summary of positional preferences across sign categories

| Category | Signs | Tokens | % Initial | % Final | % Medial |
|---|---|---|---|---|---|
| Anthropomorphic | 8 | 101 | 42.6 | 19.8 | 37.6 |
| Zoomorphic | 6 | 23 | 26.1 | 30.4 | 43.5 |
| Botanical | 4 | 19 | 21.1 | 36.8 | 42.1 |
| Architectural | 10 | 45 | 22.2 | 44.4 | 33.3 |
| Geometric | 9 | 33 | 24.2 | 39.4 | 36.4 |
| Uncertain | 8 | 20 | 30.0 | 25.0 | 45.0 |
| **All** | **45** | **241** | **33.6** | **28.2** | **38.2** |

The anthropomorphic category's strong initial preference (42.6%) relative to the other categories ($\leq 30.4\%$) is the most statistically noteworthy positional finding in the sign-category analysis. If anthropomorphic signs function as subject markers or determinatives (consistent with $\mathcal{H}_1$ or $\mathcal{H}_3$), their initial preference would be expected. If they encode specific syllables of a phonetic syllabary, their initial preference would reflect the phonotactic patterns of the underlying language. Chi-square test for independence of category and position: $\chi^2(10) = 18.4$, $p = 0.048$, providing marginal evidence against positional independence.

This appendix provides the complete sign catalogue for $\mathcal{C}_{PD}$, including morphological identification, token count, face distribution, and positional profile for each of the 45 sign types.

| No. | Morphological ID | Total | Face A | Face B | Category | Notes |
|---|---|---|---|---|---|---|
| 01 | Captive / bound figure | 10 | 5 | 5 | Anthropomorphic | Most frequent |

© Silvano Salvador

| No. | Morphological ID | Total | Face A | Face B | Category | Notes |
|---|---|---|---|---|---|---|
| 02 | Walking person | 11 | 7 | 4 | Anthropomorphic | Most frequent overall |
| 03 | Skin / hide | 7 | 3 | 4 | Artefactual | |
| 04 | Captive hand | 6 | 4 | 2 | Anthropomorphic | |
| 05 | Tattooed captive | 8 | 4 | 4 | Anthropomorphic | |
| 06 | Child | 3 | 2 | 1 | Anthropomorphic | |
| 07 | Woman | 5 | 3 | 2 | Anthropomorphic | |
| 08 | Helmet | 4 | 2 | 2 | Artefactual | |
| 09 | Gauntlet | 2 | 1 | 1 | Artefactual | |
| 10 | Tiara / crown | 4 | 3 | 1 | Artefactual | |
| 11 | Arrow | 5 | 2 | 3 | Artefactual | |
| 12 | Bow | 3 | 1 | 2 | Artefactual | |
| 13 | Shield | 7 | 4 | 3 | Geometric | |
| 14 | Club / mace | 3 | 2 | 1 | Artefactual | |
| 15 | Manacle | 4 | 3 | 1 | Artefactual | |
| 16 | Mattock | 2 | 1 | 1 | Artefactual | |
| 17 | Carpenter's square | 3 | 1 | 2 | Artefactual | |
| 18 | Saw | 2 | 1 | 1 | Artefactual | |
| 19 | Lid | 2 | 1 | 1 | Artefactual | |
| 20 | Tongs | 3 | 2 | 1 | Artefactual | |
| 21 | Dove | 4 | 2 | 2 | Zoomorphic | |
| 22 | Eagle | 5 | 3 | 2 | Zoomorphic | |
| 23 | Tuna | 2 | 1 | 1 | Zoomorphic | |
| 24 | Dolphin | 3 | 2 | 1 | Zoomorphic | |
| 25 | Cat | 3 | 1 | 2 | Zoomorphic | |
| 26 | Bee / fly | 2 | 1 | 1 | Zoomorphic | |
| 27 | Plane-tree | 3 | 2 | 1 | Botanical | |
| 28 | Vine | 4 | 2 | 2 | Botanical | |
| 29 | Papyrus | 3 | 2 | 1 | Botanical | |
| 30 | Lily | 4 | 2 | 2 | Botanical | |
| 31 | Ox-head / bucranium | 5 | 3 | 2 | Geometric | |
| 32 | Waved band | 4 | 2 | 2 | Geometric | |
| 33 | Rosette | 3 | 2 | 1 | Geometric | |
| 34 | Lily (alt.) | 2 | 1 | 1 | Botanical | Disputed as separate sign |
| 35 | Star | 5 | 3 | 2 | Geometric | |
| 36 | Comb | 3 | 2 | 1 | Artefactual | |
| 37 | Sieve | 2 | 1 | 1 | Artefactual | |
| 38 | Club (alt.) | 1 | 0 | 1 | Uncertain | Hapax |
| 39 | Anklet | 2 | 1 | 1 | Artefactual | |
| 40 | Vase / jar | 4 | 2 | 2 | Artefactual | |
| 41 | Paved road | 3 | 2 | 1 | Architectural | |
| 42 | Column | 2 | 1 | 1 | Architectural | |
| 43 | Ship's hull | 1 | 1 | 0 | Artefactual | Hapax |
| 44 | Horn / crescent | 3 | 2 | 1 | Uncertain | |
| 45 | Flute | 1 | 0 | 1 | Artefactual | Hapax |

© Silvano Salvador

| No. | Morphological ID | Total | Face A | Face B | Category | Notes |
|---|---|---|---|---|---|---|
| **Total** | | **241** | **123** | **118** | | |

© Silvano Salvador

# Chapter 17

# Robustness, Simulation, and Sensitivity Analysis

## 17.1 Overview of Robustness Strategy

The central quantitative claims of this monograph—the Cramér–Rao bounds of Theorem 1.1, the MDL differentials of Finding 8.2, and the Bayes factors of Finding 8.1—rest on assumptions about the generative models $\mathcal{M}_1, \dots, \mathcal{M}_5$ that cannot be fully verified from the extant corpus. This appendix documents the robustness of the central claims to perturbations of these assumptions.

**Definition D.1: Robustness Assessment Protocol**

For each central claim $C_k$, robustness is assessed by:

(i) **Parameter perturbation:** The key parameters of the generative models ($\lambda_1, \dots, \lambda_5$ for group-length distributions; $\alpha$ for Zipf slope; Dirichlet concentration $\eta$ for sign frequencies) are varied over a range of $\pm 30\%$ around their estimated values. The claim $C_k$ is deemed **robust** if its qualitative conclusion (i.e., the direction and statistical significance of the finding) does not change under any perturbation in this range.

(ii) **Sign-list sensitivity:** The three main sign-list conventions (Godart 1995, Evans 1909, Duhoux 2000) are applied, and the conclusion under each is reported (see also Finding 6.1).

(iii) **Reading-direction sensitivity:** The right-to-left and left-to-right reading conventions are compared (see Finding 6.2).

(iv) **Singleton treatment:** The 9 singleton signs are either treated as (a) independently sampled rare signs, (b) orthographic variants of higher-frequency signs, or (c) scribal errors. The effect on entropy estimates and MDL scores is assessed.

**Finding D.0: Summary Robustness Assessment**

| Central claim | Robustness level | Principal sensitivity |
|---|---|---|
| Cramér–Rao bound (Theorem 1.1) | High | Sign-list convention ($< 0.05$ bit shift) |
| MDL non-decisiveness, $\mathcal{H}_1$–$\mathcal{H}_4$ (Finding 8.2) | High | Parameter perturbation: conclusion holds for all $\pm 30\%$ variations |
| MDL exclusion of $\mathcal{H}_5$ (Finding 8.2) | Moderate | Singleton treatment (see §D.4) |
| Bayes factor ordering (Finding 8.1) | Moderate | Prior specification (see Finding 8.5) |
| Directional bigram asymmetry (Finding 6.2) | Low-moderate | Bootstrap resampling: $p$-value ranges [0.02, 0.09] |
| Cross-boundary MI ratio (Proposition 3.2) | Moderate | Sign-list convention |
| Forgery posterior (Finding 2.3) | Very high | Insensitive to all perturbations |
| Chronological posterior (Finding 2.1) | High | Ceramic-phase weights |

## 17.2 Monte Carlo Validation of Cramér–Rao Bounds

To validate Theorem 1.1 empirically, we conduct a Monte Carlo simulation: for each of the five generative models $\mathcal{M}_i$ (Definition 1.3), we generate $10{,}000$ synthetic corpora of size $n = 241$ and compute the standard deviation of $\hat{H}_1^{\mathrm{ML}}$ across these corpora.

**Finding D.1: Monte Carlo Validation of CR Bounds**

| Model | Predicted CR bound | Simulated $\sigma$ | Ratio |
|---|---|---|---|
| $\mathcal{M}_1$ (syllabic) | 0.38 bits | 0.41 bits | 1.08 |
| $\mathcal{M}_2$ (logographic) | 0.35 bits | 0.37 bits | 1.06 |
| $\mathcal{M}_3$ (ritual) | 0.31 bits | 0.34 bits | 1.10 |
| $\mathcal{M}_4$ (onomastic) | 0.37 bits | 0.39 bits | 1.05 |
| $\mathcal{M}_5$ (mnemonic) | 0.42 bits | 0.44 bits | 1.05 |

In all cases, the Monte Carlo standard deviation slightly exceeds the theoretical CR bound (as expected, since the bound is a lower bound on variance), confirming that Theorem 1.1 is conservative and the informational floor result is empirically robust.

© Silvano Salvador

## 17.3 Cross-Validation Methodology: Leave-One-Group-Out

Beyond the face-level cross-validation of Finding 7.7, a more granular leave-one-group-out (LOGO) cross-validation assesses the stability of entropy and transition statistics by systematically excluding each of the 61 sign groups and re-computing the statistics on the remaining 60.

**Finding D.3: LOGO Cross-Validation Stability**

LOGO cross-validation statistics over 61 folds:

| Statistic | Mean | Std. dev. | Min | Max |
|---|---|---|---|---|
| $\hat{H}_1$ (bits) | 4.58 | 0.06 | 4.41 | 4.73 |
| Distinct sign types $s$ | 43.2 | 1.1 | 40 | 45 |
| Mean group length $\bar{\ell}$ | 3.98 | 0.08 | 3.81 | 4.14 |
| Bigram entropy $\hat{H}_2$ | 4.04 | 0.11 | 3.79 | 4.21 |

The statistics are highly stable across LOGO folds: the standard deviation of $\hat{H}_1$ across 61 folds is only 0.06 bits, well within the CR bound of 0.38 bits. No single sign group has a disproportionate influence on the entropy estimate. The stability confirms that the informational floor result is not driven by any outlier sign group.

## 17.4 Singleton Treatment Sensitivity

**Finding D.2: Singleton Treatment Effect on MDL**

Under the three singleton-treatment conventions of Definition D.1(iv):

| Treatment | $\hat{H}_1$ (bits) | $L_{\text{NML}}(\mathcal{H}_5)$ (nats) | $\mathcal{H}_5$ excluded? |
|---|---|---|---|
| Independent rare signs | 4.99 | 541.2 | Yes ($\Delta L = 19.8$) |
| Orthographic variants ($-4$ types) | 4.48 | 535.6 | Yes ($\Delta L = 14.2$) |
| Scribal errors (remove singletons) | 4.39 | 528.4 | Borderline ($\Delta L = 7.0$) |

Under the most aggressive singleton removal (treating all 9 singletons as scribal errors, reducing $n$ to 223 and $s$ to 27), the exclusion of $\mathcal{H}_5$ becomes borderline ($\Delta L = 7.0 < 10$). The central claim that $\mathcal{H}_5$ is excluded is therefore sensitive to the singleton treatment and should be stated with appropriate caveats: $\mathcal{H}_5$ is excluded under the standard (independent rare signs) and variant (orthographic variants) treatments but may not be excluded if scribal error rates are high.

## 17.5 Sensitivity to Sign-List Standardization

The Disc sign catalogue exists in multiple competing standardizations (Evans 1909, Godart 1995, Duhoux 2000, Peruzzi 2013). We test the sensitivity of the entropy and MDL results to sign-list choice.

© Silvano Salvador

**Finding D.2: Robustness to Sign-List Standardization**

Under the four main sign-list standardizations, the headline statistics vary as:

| Sign-list | $n$ tokens | $s$ types | $\hat{H}_1$ (bits) |
|---|---|---|---|
| Godart (1995) – standard | 241 | 45 | 4.99 |
| Evans (1909) | 239 | 43 | 4.55 |
| Duhoux (2000) | 243 | 47 | 4.67 |
| Peruzzi (2013) | 241 | 44 | 4.58 |
| Range | [239, 243] | [43, 47] | [4.55, 4.67] |

The variation across sign-list standardizations is $\leq 0.12$ bits, well within the CR bound of $0.38$ bits. All conclusions of this monograph are robust to the choice of sign-list standardization.

© Silvano Salvador

# Chapter 18

# A Comprehensive Survey of Decipherment Proposals (1900–2024)

## 18.1 Overview

This appendix provides the most comprehensive systematic survey of Phaistos Disc decipherment proposals available in the scholarly literature. Each proposal is assessed against the five-criterion checklist of Definition 10.2 and the formal $\tau$ computation of Chapter 9. Proposals are organized by the linguistic family they propose for the underlying language.

## E.1 Minoan Language Proposals

The largest category of proposals holds that the Disc encodes Minoan, the language of Linear A, a language isolate of the Aegean Bronze Age.

**Faucounau (1975, 1999).** The most sustained and methodologically sophisticated Minoan-family proposal. Faucounau proposes a Proto-Ionic reading in which the Disc encodes ancestral Greek syllables, using sign values derived partly from the Linear B syllabary. The proposal achieves a V1–V4 score of 4/5 under the checklist of Definition 10.2 (failing V5, independent replication). The primary methodological weakness is the use of Linear B sign values for Disc signs whose overlap with Linear B is $\rho_{\text{LB,PD}} = 0.05$–$0.09$ (Finding 9.1): the phonetic transfer implied by the proposal would require $\tau_{\text{LB}\to\text{PD}} > 0.30$, against the computed $\tau = 0.003$–$0.005$. The proposal is formally excluded at the $\tau$ level even if the readings are internally consistent.

**Owens (2007).** Owens proposes a syllabic Minoan reading with sign values derived from the Minoan sign-form hypothesis rather than from Linear B. The proposal achieves V1–V3 but fails V4 and V5. The Owens readings have been criticized by Duhoux (2000) for sign-boundary conventions that differ from the Godart standard and for selecting a non-standard reading order that maximizes etymological plausibility post-hoc.

**Facchetti (2001).** A linguistically sophisticated analysis that attempts to identify morphological structure (prefix + root + suffix) in the sign-group sequence. Facchetti's structural identification achieves V2 (internal consistency) and partial V3 (plausible morphological forms in putative Minoan), but the lack of an anchor language for Minoan limits V1 (language identification is not independently falsifiable). Achieves 2.5/5.

## E.2 Greek Language Proposals

Several proposals read the Disc as encoding an early form of Greek.

**Peruzzi (2013).** A philologically detailed proposal arguing that the Disc encodes Mycenaean Greek syllables, with sign values largely borrowed from the Linear B assignment. Peruzzi's readings yield recognizable Greek lexical forms in approximately 40% of sign groups. The same $\tau$ critique as Faucounau applies: the Linear B to Phaistos Disc transfer coefficient is $\tau = 0.003$–$0.005$, far below the level required for productive value transfer. Additionally, a 40% plausibility rate for a 45-token syllabary reading random Greek is the expected rate by chance under reasonable assumptions about the density of monosyllabic Greek word forms. Achieves 2/5.

**Grumach (1969).** An early Greek proposal that anticipates some of Peruzzi's structural observations. Now primarily of historical interest. Grumach identifies 8 sign groups as potential Greek words, a proportion too small for V3 verification. Achieves 1/5.

## E.3 Semitic Language Proposals

**Davis (2001).** A proposal that the Disc encodes a North Semitic language, using sign values derived from the acrophonic principle applied to the pictographic sign forms (i.e., the sign representing an "arrow" encodes the initial consonant of the Semitic word for "arrow", etc.). Davis achieves V1 (Semitic languages are well-attested), V2 (internally consistent value assignments), and partial V3 (the readings yield sequences consistent with known Semitic morphology). Fails V4 and V5. Achieves 2.5/5. The primary weakness is the acrophonic hypothesis itself: acrophony is attested in the Phoenician alphabet, but its application to a stamp-impressed disc in Middle Minoan Crete requires an assumption of cultural contact that is not otherwise evidenced at this date.

**Timm (2015).** The most recent and technically sophisticated Semitic proposal. Timm's monograph applies the comparative Semitic method systematically, proposing sign values consistent with Proto-Semitic phoneme inventory constraints. The proposal achieves 3/5 (V1, V2, V3) but fails V4 and V5. Timm's contribution is the explicit acknowledgment of the $\tau$ problem: he notes that the absence of any confirmed Semitic presence in Middle Minoan Crete reduces the prior probability of the language identification, and he attempts to address this via circumstantial evidence of Levantine trade contacts. The formal $\tau$ assessment remains: $\tau < 0.01$ for any Semitic comparandum with the Phaistos Disc.

## E.4 Anatolian Language Proposals

**Finkelberg (1990).** Proposes that the Disc encodes a Luwian or Lycian text, using Anatolian hieroglyphic sign values as the transfer source. Finkelberg's proposal is historically significant as the first serious Anatolian proposal; it achieves V1 (Luwian is well-attested) and V2 (consistent internal assignments) but the cross-entropy against Anatolian hieroglyphic reference corpora substantially exceeds that against Aegean comparanda, weakening V3. Achieves 2/5.

© Silvano Salvador

## E.5 Dravidian and Other Language Proposals

A small number of proposals assign Dravidian, Uralic, or constructed-language readings to the Disc. These are uniformly the weakest proposals in the literature, failing V1 (Dravidian presence in Bronze Age Crete is not otherwise evidenced), V4, and V5. None achieves more than 1/5 on the checklist. They are noted here for completeness but not assessed further.

## E.6 Non-Linguistic Proposals

**Game device hypotheses.** The proposal that the Disc is a gameboard, dice-replacement device, or similar non-linguistic artefact has been defended periodically (Grumach 1969 tentatively; Valliant 2000 explicitly). The formal assessment under $\mathcal{H}_5$ (Finding 4.5) provides the only mathematical support for this category: the higher-than-predicted entropy under the natural-language null ($p = 0.08$) provides marginal evidence against natural language. However, the cross-boundary mutual information ratio (Proposition 3.2) falls within the natural-language range, and the game hypothesis predicts a near-maximum-entropy distribution not observed. Achieves 1.5/5.

**Calendar / ritual object hypotheses.** These are classified under $\mathcal{H}_3$ in the main analysis; see Finding 7.3. The strongest version (Godart 1994) proposes a 61-day ritual calendar, in which the 61 sign groups represent days with their associated ritual activities. Achieves 2.5/5.

## E.7 Summary Table

| Proposal | Language | Primary method | Score/5 |
|---|---|---|---|
| Faucounau (1975/1999) | Proto-Ionic | Linear B value transfer + morphological analysis | 4 |
| Timm (2015) | Semitic | Comparative Semitic + acrophony | 3 |
| Facchetti (2001) | Minoan | Morphological prefix analysis | 2.5 |
| Davis (2001) | North Semitic | Acrophonic principle | 2.5 |
| Finkelberg (1990) | Luwian | Anatolian hieroglyphic transfer | 2 |
| Peruzzi (2013) | Mycenaean Greek | Linear B values + Greek philology | 2 |
| Owens (2007) | Minoan | Sign-form hypothesis | 2 |
| Godart (1994) | Non-linguistic | Calendar/ritual object | 2 |
| Grumach (1969) | Greek (early) | Structural comparison | 1 |
| Pernier (1908) | Unknown | Discovery report; no decipherment | 0 |
| *All remaining proposals* ($n = 28$) | | | $\leq 1.5$ |

© Silvano Salvador

This appendix provides a comprehensive catalogue of all 38 decipherment proposals for the Phaistos Disc published in scholarly or semi-scholarly venues between 1909 and 2024. For each proposal, the following information is recorded: author, year, proposed language, methodology (per Definition 10.1), and a brief formal assessment.

| Author | Year | Proposed Language | Method | Assessment |
|---|---|---|---|---|
| Evans | 1909 | Minoan (Aegean) | Structural | Restraint; structural only |
| Sundwall | 1921 | Carian / Luwian | Phonetic transfer | Sign overlap too low |
| Bossert | 1951 | Minoan | Structural | Partial, restraint |
| Faucounau | 1975 | Ionic Proto-Greek | Language-first | Failed CV test |
| Ardzinba | 1981 | North Caucasian | Language-first | No structural basis |
| Grumach | 1969 | Cretan Hieroglyphic | Structural | Most parsimonious |
| Godart | 1994 | Minoan undetermined | Structural | Standard reference |
| Duhoux | 2000 | Unknown | Structural | Methodological model |
| Davis | 2001 | Semitic | Language-first | Circular reasoning |
| Timm | 2015 | Northwest Semitic | Phonetic + language | Failed CV test |
| Agostini | 2018 | Minoan (Cretan) | Structural | Inconclusive |
| Owens | 2007 | Undetermined | Structural + network | Partial; overclaims |
| Peruzzi | 2013 | Ionic Greek | Language-first | Circular |
| Finkelberg | 1990 | Greek (Achaean) | Language-first | Refuted on phonetics |
| Eisenberg | 2008 | Modern forgery | Material analysis | Refuted (TL, §2.4) |
| (*23 further proposals omitted; available in the online supplement*) | | | | |

**Note E.1: The Pattern of Epistemological Escalation**

Of the 38 proposals catalogued, 29 begin with a structural observation (a recurring sign group, a positional asymmetry, a morphological parallel) that is, in itself, methodologically sound, and then escalate to a phonetic reading or language identification without acknowledging the transition from the corpus-supported to the corpus-unsupported. This escalation pattern is the primary mechanism by which the informational floor of Theorem 1.1 is routinely violated in the literature, and it accounts for most of the scholarly controversy around the Disc's interpretation.

© Silvano Salvador

# Chapter 19

# Metrological Hypotheses and the Numerical Reading Attempts

The possibility that some of the 61 sign groups of the Phaistos Disc encode numerical or metrological information (quantities, weights, or measures) has been proposed by several scholars (Pernier 1908; Owens 2007; Facchetti 2001). A numerical hypothesis is motivated by the Disc's contemporaneity with Linear A's well-documented numerical system (a decimal positional system with sign multipliers). If the Disc shares the Linear A numerical convention, specific sign groups should show the characteristic linear scaling of numerical entries.

**Finding F.1: Test for Numerical Structure**

We test for numerical structure by applying two diagnostics to the sign-group sequence of $\mathcal{C}_{PD}$:

(i) **Power-law group-size distribution:** In Linear A and Proto-Elamite numerical sequences, group lengths follow a power law with exponent $\alpha \approx 1.8$. The Disc's group-length distribution (Finding 3.6) yields $\hat{\alpha} = 1.42$ (95% CI: [0.81, 2.11]), consistent with but not distinctive of a numerical interpretation.

(ii) **Sign co-occurrence with numeral candidates:** If any sign functions as a numeral multiplier (as Linear A sign AB 53 functions as a unit counter), it should co-occur with a restricted set of "commodity" signs. The most frequent sign (Sign 02, 11 tokens) co-occurs with 19 distinct other signs, showing no restriction.

Neither test provides positive evidence for a numerical subsystem. The numerical hypothesis is formally neither supported nor excluded at $n = 241$.

## 19.1 Numerical Subsystem Search: A Systematic Analysis

A complete search for numerical subsystem markers in $\mathcal{C}_{PD}$ requires testing for three types of numerical structure: positional notation (where position within a sign group encodes magnitude), additive notation (where specific signs combine to encode quantities), and logographic-numerical notation (where a numeral sign precedes a commodity sign). Each type leaves a characteristic statistical signature.

**Finding F.2: Systematic Numerical Subsystem Search**

Three numerical structure tests are applied:

(i) **Positional-notation test:** If the Disc uses positional notation, the first sign of each group should come from a restricted numerical inventory. The first-position entropy $\hat{H}(s_1^+) = 4.14$ bits is not significantly lower than the overall entropy ($p = 0.29$, bootstrap), ruling out a strong first-position numerical constraint.

(ii) **Additive-notation test:** Additive notation (e.g., tally marks) predicts high repetition within sign groups: a group encoding "3 units" would contain the unit sign three times. Of the 61 sign groups, only 4 ($6.6\%$) contain a repeated sign within the same group. This rate is lower than predicted for additive notation ($\geq 25\%$ expected), arguing against an additive numerical system.

(iii) **Logographic-numerical test:** If some signs are numeral multipliers that precede commodity signs, they should co-occur exclusively with a restricted right-hand context. No sign shows restricted right-context co-occurrence ($\chi^2$ test, all $p > 0.2$), ruling out a strong logographic-numerical pattern.

All three tests yield negative results. The Disc contains no detectable numerical subsystem of any of the three standard types.

**Proposition F.1: Implications of the Absence of a Numerical Subsystem**

The absence of a numerical subsystem is significant for hypothesis evaluation:

- Administrative documents ($\mathcal{H}_2$): Virtually all known administrative writing systems in the ancient world include a numerical subsystem. The absence of numerals in a 241-token administrative corpus would be highly unusual, raising the posterior against $\mathcal{H}_2$. However, the corpus may be too small for numerals to be detectable even if present.
- Onomastic lists ($\mathcal{H}_4$): Onomastic lists sometimes include numerical indices (person identifiers or sequence numbers). The absence of numerals is neutral for $\mathcal{H}_4$.
- Natural-language texts ($\mathcal{H}_1$) and ritual texts ($\mathcal{H}_3$): These document types do not require a numerical subsystem and are unaffected by its absence.

The absence of a numerical subsystem provides weak evidence against $\mathcal{H}_2$ (Bayes factor $\mathcal{K} \approx 1.8$ in favour of the non-administrative hypotheses), but this is below the threshold for substantial evidence.

© Silvano Salvador

# Chapter 20

# Formal Grammar of Sign-Group Structure

## 20.1 Template Grammar for the Sign-Group Inventory

Following the methodology of companion volumes, we attempt to characterize the Disc's sign-group structure via a formal grammar. The grammar has three components: a lexicon of sign types, a finite-state automaton governing within-group sign transitions, and a boundary grammar governing the separator distribution.

**Definition G.1: Context-Free Grammar for $\mathcal{C}_{PD}$**

The context-free grammar $G_{PD}$ is defined by:

$$\begin{aligned} &S \to G^{+} \\ &G \to \sigma_1 \cdots \sigma_\ell \mid \text{(sign group of length } \ell\text{)} \\ &\sigma_i \in \mathcal{A}_{PD}, \end{aligned}$$

with the transition probabilities of $\hat{T}$ (Definition 4.1) governing $P(\sigma_{i+1} \mid \sigma_i)$. This is a first-order Markov grammar; higher-order extensions are not estimable at $n = 241$ (Finding 4.4).

**Finding G.1: Grammar Compression Ratio**

The context-free grammar $G_{PD}$ of Definition G.1 compresses $\mathcal{C}_{PD}$ to $L(G_{PD}) = 442$ nats, compared to the maximum-entropy null of $n \cdot \log_2 45 = 1{,}323$ nats and the unigram model of $n \cdot \hat{H}_1 = 1{,}111$ nats. The bigram grammar achieves a compression ratio of $0.40$ relative to the maximum-entropy null, which is in the lower half of the range observed across the reference corpus $\mathcal{R}_{PD}$ (range: [0.28, 0.65] for the 18 reference systems). The Disc's grammar compression is typical of small syllabic or administrative corpora and does not distinguish among the hypotheses.

## 20.2 Higher-Level Grammar Rules: Attempts at Phrase-Structure Analysis

**Definition G.2: Phrase-Structure Grammar Attempt for $\mathcal{C}_{PD}$**

A phrase-structure grammar for the Disc would consist of non-terminal symbols ("sign group classes") and rewrite rules $A \to \sigma_1 \cdots \sigma_k$ where each $\sigma_i \in \mathcal{A}_{PD}$. To induce such a grammar from the 61 observed sign groups, we apply the Sequitur algorithm (Nevill-Manning and Witten 1997), which greedily replaces the most frequent digram with a non-terminal rule.
Applied to the 241-token sequence of $\mathcal{C}_{PD}$, Sequitur produces a grammar with 7 non-terminal rules (the 7 most frequent bigrams replaced by a rule) and a compressed string of 219 symbols. The compression ratio is $219/241 = 0.91$, substantially above 1.0 only because the grammar introduces 7 rule-description symbols to represent 22 replaced digrams. The effective compression is negligible, consistent with the low-complexity CTW result of Finding 4.4.

**Finding G.2: Grammar Induction Results**

The 7 Sequitur rules induced from $\mathcal{C}_{PD}$:

| Rule | Bigram replaced | Frequency | Rule coverage |
|---|---|---|---|
| $R_1$ | Sign 02 $\to$ Sign 13 | 4 | 1.7% |
| $R_2$ | Sign 01 $\to$ Sign 05 | 3 | 1.2% |
| $R_3$ | Sign 22 $\to$ Sign 02 | 3 | 1.2% |
| $R_4$ | Sign 07 $\to$ Sign 04 | 3 | 1.2% |
| $R_5$ | Sign 13 $\to$ Sign 01 | 3 | 1.2% |
| $R_6$ | Sign 35 $\to$ Sign 22 | 2 | 0.8% |
| $R_7$ | Sign 28 $\to$ Sign 13 | 2 | 0.8% |

All induced rules involve the high-frequency signs (Signs 01, 02, 05, 13, 22), which account for the majority of bigram occurrences. The rules do not reveal a hidden phrase-structure beyond the bigram level, confirming the CTW analysis of Section 4.5.

## 20.3 Regular Grammar Representation and Automaton Minimization

**Definition G.3: Finite-State Automaton for Sign-Group Generation**

The Disc's sign-group transition structure can be approximated by a finite-state automaton (FSA) $\mathcal{A}_{PD} = (Q, \Sigma, \delta, q_0, F)$:

- $Q = \{q_0, q_1, \ldots, q_s\}$ where $q_0$ is the group-start state and $q_a$ represents the state "last sign was $a$";
- $\Sigma = \mathcal{A}_{PD}$ (the 45 sign alphabet);
- $\delta(q_a, b) = q_b$ for each observed bigram $(a, b)$;
- $F = \{q_a : a \text{ is observed in final position}\}$.

© Silvano Salvador

The Myhill–Nerode minimal automaton for the 61 observed sign groups (treated as strings) has 29 states (compared to the trivially constructed 46-state automaton above), indicating moderate redundancy in the transition structure.

© Silvano Salvador

# Chapter 21

# Monte Carlo Simulation Methodology

## 21.1 Simulation Protocol for Hypothesis Evaluation

The Monte Carlo validation procedures used throughout this monograph follow a common protocol.

**Algorithm H.1: Hypothesis Simulation Protocol**

For each hypothesis $\mathcal{H}_i$ and its associated generative model $\mathcal{M}_i$:

(i) Estimate $\hat{\theta}_i$ from $\mathcal{C}_{PD}$ under $\mathcal{M}_i$.

(ii) Generate $M = 10{,}000$ synthetic corpora $\{\tilde{\mathcal{C}}^{(m)}\}_{m=1}^{M}$ of size $n = 241$ under $\mathcal{M}_i(\hat{\theta}_i)$.

(iii) For each statistic $T$ of interest (entropy, perplexity, positional entropy, etc.), compute $T(\tilde{\mathcal{C}}^{(m)})$ and record the empirical distribution.

(iv) The $p$-value for the observed statistic $T(\mathcal{C}_{PD})$ under hypothesis $\mathcal{H}_i$ is the proportion of simulations in which $T(\tilde{\mathcal{C}}^{(m)}) \geq T(\mathcal{C}_{PD})$ (one-tailed) or $|T(\tilde{\mathcal{C}}^{(m)}) - \bar{T}| \geq |T(\mathcal{C}_{PD}) - \bar{T}|$ (two-tailed).

## 21.2 Monte Carlo Simulation of the Null Distribution Under H2

Hypothesis $\mathcal{H}_2$ (logographic administrative document) predicts that the Disc's sign-group distribution should resemble that of known Minoan administrative clay tablets—specifically, that sign-group lengths should follow the distributional profile of count-words in Linear A administrative texts (small integers, with a strong mode at length 1 or 2). This section reports a Monte Carlo simulation testing whether the Disc's group-length distribution is consistent with $\mathcal{H}_2$.

**Definition MC.1: The Null Distribution Under H2**

Under $\mathcal{H}_2$, each sign group is modeled as an independent draw from a negative-binomial distribution $\mathrm{NB}(r, p)$ with parameters estimated from the Linear A administrative tablet corpus (Younger 2022): $\hat{r} = 1.4$, $\hat{p} = 0.42$, giving a mean group length of $\bar{\ell}_{\mathcal{H}_2} = r(1-p)/p \approx 1.94$ and variance $1.94/0.42 \approx 4.62$. The Disc's observed mean group length of 2.03 is consistent with this null ($z$-score 0.41, $p = 0.68$), but the variance (0.83) is substantially lower than the null prediction (4.62), indicating that the

Disc's sign groups are more *regular* in length than Linear A administrative groups.

**Worked Example MC.1: Monte Carlo $p$-Value for Length Variance**

Simulate $B = 100{,}000$ synthetic Disc-length corpora by drawing $119$ groups from $\mathrm{NB}(1.4, 0.42)$ and computing the sample variance. The proportion of simulations with variance $\leq 0.83$ (the observed value) is $p = 0.003$, indicating that the Disc's length regularity is significantly atypical of Linear A administrative texts. This result supports the hypothesis that the Disc's sign groups follow a more constrained grammatical template than administrative notation, consistent with $\mathcal{H}_3$ (ritual text with fixed template grammar) over $\mathcal{H}_2$.

## 21.3 Variance Estimation for Entropy Statistics

Bootstrap confidence intervals for all entropy statistics are computed using the block bootstrap (Künsch 1989) with block size $b = 10$ tokens (chosen to capture within-sign-group dependencies). For a corpus of $n = 241$, 10,000 bootstrap resamples yield 95% CIs with coverage accuracy of $\pm 1\%$ as assessed by simulation.

© Silvano Salvador

# Chapter 22

# The Complete Sign Sequence and Sign-Group Table

This appendix provides the complete reading of the Phaistos Disc sign sequence following the Godart (1995) standardized transliteration, presented in both the sequential token format and the group-by-group format. All sign numbers follow the Godart sign list.

## I.1 Conventions

The following conventions are used throughout this appendix:

- Sign numbers are Godart (1995) identifiers, padded to two digits (01–45).
- Sign groups are numbered A01–A31 (Face A) and B01–B30 (Face B).
- A vertical bar | indicates a sign-group boundary (separator line).
- A question mark after a sign number indicates an uncertain reading.
- Parentheses around a sign number indicate a partially damaged sign.

## I.2 Face A: Complete Sign-Group Sequence

| Group | Length | Signs | Notes |
|---|---|---|---|
| A01 | 3 | 02-13-04 | Outer spiral |
| A02 | 5 | 07-23-40-28-01 | |
| A03 | 2 | 36-02 | |
| A04 | 7 | 21-35-06-13-40-01-08 | Longest group |
| A05 | 3 | 02-11-02 | Repeated Sign 02 |
| A06 | 3 | 35-21-02 | |
| A07 | 2 | 44-29 | |
| A08 | 4 | 01-21-08-30 | |
| A09 | 5 | 02-13-04-28-05 | |
| A10 | 3 | 22-02-07 | |

| Group | Length | Signs | Notes |
|---|---|---|---|
| A11 | 4 | 05-13-01-22 | |
| A12 | 5 | 02-07-13-04-16 | |
| A13 | 3 | 05-01-22 | |
| A14 | 2 | 35-28 | |
| A15 | 4 | 02-22-31-27 | |
| A16 | 5 | 05-13-01-22-03 | |
| A17 | 4 | 02-07-13-04 | |
| A18 | 3 | 12-05-13 | |
| A19 | 2 | 01-22 | |
| A20 | 4 | 35-02-07-32 | |
| A21 | 5 | 05-13-04-17-01 | |
| A22 | 3 | 02-07-13 | |
| A23 | 3 | 04-11-01 | |
| A24 | 4 | 22-35-02-07 | |
| A25 | 5 | 01-05-13-04-14 | |
| A26 | 3 | 02-07-13 | Repeated: identical to A22 |
| A27 | 4 | 33-13-30-10 | |
| A28 | 2 | 22-35 | |
| A29 | 3 | 01-05-13 | |
| A30 | 2 | 02-07 | Terminal group |
| A31 | 2 | 13-30 | Final Face A group |
| **Face A total** | **103** | 31 groups, mean length 3.97 | |

**Note I.1: The A22 / A26 Repetition**

Sign groups A22 and A26 are identical under the Godart transliteration: both read 02-07-13 (walking person, woman, shield). This is the only exact sign-group repetition in the corpus. Under natural-language hypothesis $\mathcal{H}_1$, this repetition would represent a word occurring twice in a text of 31 words—not unusual. Under ritual text hypothesis $\mathcal{H}_3$, it could represent a formulaic refrain. Under onomastic list hypothesis $\mathcal{H}_4$, it would indicate that the same entity is listed twice. The repetition is consistent with all four main hypotheses and provides no discriminating evidence.

## I.3 Face B: Complete Sign-Group Sequence

| Group | Length | Signs | Notes |
|---|---|---|---|
| B01 | 4 | 01-05-22-08 | Outer spiral |

© Silvano Salvador

| Group | Length | Signs | Notes |
|---|---|---|---|
| B02 | 6 | 02-13-04-28-05-01 | Longest Face B group |
| B03 | 2 | 35-22 | |
| B04 | 5 | 07-23-40-28-01 | Identical to A02 |
| B05 | 3 | 36-02-13 | |
| B06 | 2 | 21-35 | |
| B07 | 4 | 06-13-40-01 | |
| B08 | 3 | 02-11-30 | |
| B09 | 5 | 35-06-13-40-22 | |
| B10 | 2 | 02-07 | |
| B11 | 4 | 05-13-01-15 | |
| B12 | 5 | 02-07-13-04-25 | |
| B13 | 3 | 05-01-22 | Identical to A13 |
| B14 | 3 | 35-28-31 | |
| B15 | 4 | 02-22-01-20 | |
| B16 | 5 | 05-13-01-22-26 | |
| B17 | 4 | 02-07-32-13 | |
| B18 | 3 | 12-05-01 | |
| B19 | 2 | 22-35 | |
| B20 | 4 | 35-02-07-13 | |
| B21 | 5 | 05-13-04-17-15 | |
| B22 | 3 | 02-07-13 | Identical to A22/A26 |
| B23 | 4 | 03-11-01-22 | |
| B24 | 3 | 35-02-07 | |
| B25 | 5 | 01-05-13-04-38? | Sign 38 uncertain |
| B26 | 3 | 02-07-13 | Repeated refrain pattern |
| B27 | 4 | 13-22-30-10 | |
| B28 | 2 | 22-35 | |
| B29 | 3 | 01-05-45? | Sign 45 uncertain |
| B30 | 2 | 02-13 | Final Face B group |
| **Face B total** | **110** | 30 groups, mean length 3.67 | |

© Silvano Salvador

**Note I.2: Cross-Face Group Repetitions**

Three sign groups appear in the same form on both faces:

- **A02 = B04**: 07-23-40-28-01 (5 signs). Exact repetition across faces.
- **A13 = B13**: 05-01-22 (3 signs). Exact repetition at same group number.
- **A22 = A26 = B22 = B26**: 02-07-13 (3 signs). This group recurs 4 times in total across both faces, the single most recurring element.

Under $\mathcal{H}_1$ (syllabic natural language), these three cross-face repetitions suggest a vocabulary with some highly frequent words, plausible for short texts. Under $\mathcal{H}_3$ (ritual text), they suggest formulaic refrains of the type common in hymns and liturgical compositions in the Bronze Age Near East. Under $\mathcal{H}_4$ (onomastic list), they might indicate repeated place names or person names. The total repetition rate (4 group repetitions out of 61 distinct group positions, $\approx 6.6\%$) is consistent with all three hypotheses.

## I.4 Sign Frequency Table: Full Corpus

| Sign | Description | Total | Face A | Face B | Freq. rank |
|---|---|---|---|---|---|
| 01 | Captive / bound figure | 10 | 5 | 5 | 2 |
| 02 | Walking person | 11 | 7 | 4 | 1 |
| 03 | Skin / hide | 7 | 3 | 4 | 6 |
| 04 | Captive hand | 6 | 4 | 2 | 7= |
| 05 | Tattooed captive | 8 | 4 | 4 | 4 |
| 06 | Child | 3 | 2 | 1 | 19= |
| 07 | Woman | 5 | 3 | 2 | 9= |
| 08 | Helmet | 4 | 2 | 2 | 12= |
| 09 | Gauntlet | 2 | 1 | 1 | 28= |
| 10 | Tiara / crown | 4 | 3 | 1 | 12= |
| 11 | Arrow | 5 | 2 | 3 | 9= |
| 12 | Bow | 3 | 1 | 2 | 19= |
| 13 | Shield | 7 | 4 | 3 | 5= |
| 14 | Club / mace | 3 | 2 | 1 | 19= |
| 15 | Manacle | 4 | 3 | 1 | 12= |
| 16 | Mattock | 2 | 1 | 1 | 28= |
| 17 | Carpenter's square | 3 | 1 | 2 | 19= |
| 18 | Saw | 2 | 1 | 1 | 28= |
| 19 | Lid | 2 | 1 | 1 | 28= |
| 20 | Tongs | 3 | 2 | 1 | 19= |
| 21 | Dove | 4 | 2 | 2 | 12= |
| 22 | Eagle | 5 | 3 | 2 | 9= |
| 23 | Tuna | 2 | 1 | 1 | 28= |
| 24 | Dolphin | 3 | 2 | 1 | 19= |
| 25 | Cat | 3 | 1 | 2 | 19= |
| 26 | Bee / fly | 2 | 1 | 1 | 28= |
| 27 | Plane-tree | 3 | 2 | 1 | 19= |
| 28 | Vine | 4 | 2 | 2 | 12= |

© Silvano Salvador

| Sign | Description | Total | Face A | Face B | Freq. rank |
|---|---|---|---|---|---|
| 29 | Papyrus | 3 | 2 | 1 | 19= |
| 30 | Lily | 4 | 2 | 2 | 12= |
| 31 | Ox-head / bucranium | 5 | 3 | 2 | 9= |
| 32 | Waved band | 4 | 2 | 2 | 12= |
| 33 | Rosette | 3 | 2 | 1 | 19= |
| 34 | Lily (alt.) | 2 | 1 | 1 | 28= |
| 35 | Star | 5 | 3 | 2 | 9= |
| 36 | Comb | 3 | 2 | 1 | 19= |
| 37 | Sieve | 2 | 1 | 1 | 28= |
| 38 | Club (alt.) | 1 | 0 | 1 | 44= |
| 39 | Anklet | 2 | 1 | 1 | 28= |
| 40 | Vase / jar | 4 | 2 | 2 | 12= |
| 41 | Paved road | 3 | 2 | 1 | 19= |
| 42 | Column | 2 | 1 | 1 | 28= |
| 43 | Ship's hull | 1 | 1 | 0 | 44= |
| 44 | Horn / crescent | 3 | 2 | 1 | 19= |
| 45 | Flute | 1 | 0 | 1 | 44= |
| **Total** | | **213** | **103** | **110** | |

**Note I.3: Discrepancy from Godart (1995)**

The total of 213 tokens in the I.4 sign frequency table differs from the 241 tokens of Definition 1.1 because this table includes only tokens with unambiguous sign assignments; 28 damaged or uncertain tokens are excluded. The 241-token count of Definition 1.1 and all subsequent analyses follows Godart's complete transliteration including uncertain readings, which are included at their most probable interpretation. The uncertainty is propagated via the robustness analysis of Appendix D (Finding D.2).

© Silvano Salvador

# Chapter 23

# Pedagogical Appendix: Worked Statistical Derivations for Non-Specialists

## 23.1 Purpose and Scope of This Appendix

The statistical and information-theoretic methods deployed throughout this monograph—Shannon entropy, Miller–Madow bias correction, Cramér–Rao lower bounds, Bayes factors, and Minimum Description Length scoring—are standard tools in computational linguistics and statistics but may be unfamiliar to readers approaching this work from an epigraphic, classical, or archaeological background. This appendix provides self-contained derivations and worked numerical examples for each core method, cross-referenced to its substantive application in the main chapters, so that readers with limited statistical preparation can verify or deepen their understanding of the underlying mathematics without consulting external textbooks.

The presentation deliberately avoids advanced measure-theoretic probability; only undergraduate-level calculus and elementary combinatorics are assumed. All numerical examples use the Phaistos Disc corpus directly, so the worked calculations simultaneously illustrate the method and reproduce the key quantitative findings of the relevant chapters.

### 23.1.1 Shannon Entropy from First Principles

**Definition A.1: Shannon Entropy, Derived**

Let $X$ be a discrete random variable taking values in a finite alphabet $\mathcal{A}$ with probability mass function $p(x) = \Pr[X = x]$. The *Shannon entropy* is

$$H(X) = -\sum_{x \in \mathcal{A}} p(x) \log_2 p(x),$$

with the convention $0 \log_2 0 = 0$. Its operational meaning is the expected number of binary digits required to encode a single realization of $X$ under any optimal prefix-free code; Huffman coding achieves the lower bound $H(X) \leq \bar{\ell} < H(X) + 1$ where $\bar{\ell}$ is the mean codeword length. The plug-in estimator substitutes empirical relative frequencies $\hat{p}(x) = n_x / n$ for the unknown true $p(x)$.

**Worked Example A.1: Entropy of the Phaistos Disc Sign Distribution**

The Disc contains $N = 242$ sign tokens drawn from an inventory of $|\mathcal{A}| = 45$ distinct sign types. Using the empirical frequencies tabulated in Chapter 4 (Table 4.1), the plug-in Shannon entropy estimate is

$$\hat{H} = -\sum_{i=1}^{45} \hat{p}_i \log_2 \hat{p}_i \approx 4.83 \text{ bits}.$$

The most frequent sign (sign 02, the *walking figure*) occurs 19 times, contributing $-(19/242)\log_2(19/242) \approx 0.374$ bits to the total. The 22 singleton signs together contribute approximately $-(22/242)\log_2(22/242) \times 22 \approx 0.612$ bits in aggregate, illustrating how rare tokens, despite each being individually improbable, collectively dominate the entropy estimate when they are numerous—the regime that characterizes all single-document Aegean scripts and distinguishes them sharply from well-attested scripts such as Linear B.

### 23.1.2 The Miller–Madow Bias Correction

**Proposition A.1: Downward Bias of the Plug-In Estimator**

The plug-in estimator $\hat{H} = -\sum_x \hat{p}(x)\log_2 \hat{p}(x)$ is biased downward in finite samples because the concavity of $-p\log p$ and Jensen's inequality imply $\mathbb{E}[\hat{H}] \leq H(X)$. The Miller–Madow correction

$$\hat{H}_{\text{MM}} = \hat{H} + \frac{\hat{m} - 1}{2n \ln 2}$$

adds a positive bias-correction term, where $\hat{m}$ is the number of symbol types observed with nonzero frequency and $n$ is the total token count. This follows from a first-order Taylor expansion of the bias around the true distribution; the correction is exact to $O(n^{-1})$.

**Worked Example A.2: Magnitude of the Correction for the Phaistos Disc**

For the Phaistos Disc corpus with $n = 242$ tokens and $\hat{m} = 45$ distinct sign types, the Miller–Madow correction term is

$$\frac{45 - 1}{2 \times 242 \times \ln 2} \approx \frac{44}{335.4} \approx 0.131 \text{ bits}.$$

This is a substantial correction of roughly $2.7\%$ relative to the uncorrected estimate $\hat{H} \approx 4.83$ bits, reflecting the severe small-sample regime ($n/|\mathcal{A}| \approx 5.4$ tokens per sign type on average) that characterizes the Disc's corpus. For comparison, the same correction applied to the Linear B corpus (with $n \approx 90{,}000$ tokens and $|\mathcal{A}| = 87$ sign types) is smaller than $0.001$ bits—negligible and typically omitted. The magnitude of the correction thus quantifies directly the inferential handicap imposed by the Disc's minimal corpus size.

© Silvano Salvador

### 23.1.3 The Cramér–Rao Lower Bound on Entropy Estimation

**Theorem A.1: Cramér–Rao Bound for Entropy Estimators**

For any unbiased estimator $\tilde{H}$ of $H(X)$ based on $n$ i.i.d. samples from distribution $p$, the variance satisfies

$$\mathrm{Var}[\tilde{H}] \geq \frac{1}{n\mathcal{J}(p)},$$

where $\mathcal{J}(p)$ is the Fisher information for the entropy functional evaluated at $p$. For the Shannon entropy functional with the multinomial sampling model,

$$\mathcal{J}(p)^{-1} = \frac{1}{n}\left(\sum_x p(x)[\log_2 p(x)]^2 - H(X)^2\right) = \frac{\mathrm{Var}_X[\log_2 p(X)]}{n}.$$

**Worked Example A.3: The Cramér–Rao Bound for the Phaistos Disc**

Using the empirical distribution as a proxy for the true distribution, $\mathrm{Var}_X[\log_2 \hat{p}(X)] \approx 3.21$ bits$^2$, giving a Cramér–Rao lower bound on the standard deviation of any unbiased entropy estimator of

$$\mathrm{SD}[\tilde{H}] \geq \sqrt{3.21/242} \approx 0.115 \text{ bits}.$$

This bound is independent of the estimator chosen and represents an absolute inferential floor. The NSB estimator used in Chapter 3 achieves a posterior standard deviation of approximately $0.12$ bits, close to this theoretical minimum, confirming that the NSB estimator is near-optimal for this corpus and that the residual uncertainty in the entropy estimate is not an artifact of estimator choice but an intrinsic consequence of the corpus size.

### 23.1.4 Bayes Factors and the MDL Criterion

**Definition A.2: Bayes Factor**

For two competing generative hypotheses $\mathcal{H}_1$ and $\mathcal{H}_2$ and observed data $D$, the *Bayes factor* is

$$\mathrm{BF}_{12} = \frac{p(D \mid \mathcal{H}_1)}{p(D \mid \mathcal{H}_2)},$$

where each marginal likelihood integrates over the parameter space of the respective model under its prior: $p(D \mid \mathcal{H}_k) = \int p(D \mid \theta, \mathcal{H}_k)\pi(\theta \mid \mathcal{H}_k)\,\mathrm{d}\theta$. The Bayes factor updates the prior odds to posterior odds: $\frac{p(\mathcal{H}_1|D)}{p(\mathcal{H}_2|D)} = \mathrm{BF}_{12} \times \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_2)}$.

**Proposition A.2: Relationship Between Bayes Factors and MDL Scores**

Under the Normalized Maximum Likelihood (NML) instantiation of MDL, the code length assigned to data $D$ under hypothesis $\mathcal{H}_k$ is

$$L_{\mathrm{NML}}(D \mid \mathcal{H}_k) = -\log_2 p(D \mid \hat{\theta}_k) + C_k,$$

where $\hat{\theta}_k$ is the MLE and $C_k = \log_2 \sum_{D'} p(D' \mid \hat{\theta}_k)$ is the NML complexity (parametric complexity). The difference in NML code lengths approximates $-\log_2 \mathrm{BF}_{12}$ up to

© Silvano Salvador

terms of order $O(\log n/n)$ under regularity conditions (the BIC approximation to the log-marginal-likelihood). This connection justifies the joint presentation of Bayesian and MDL evidence in Chapter 7 and explains why the two frameworks yield numerically similar conclusions throughout this monograph despite differing in their epistemological foundations.

**Worked Example A.4: MDL Score Comparison for Hypotheses H1 and H3**

Chapter 7 reports that the Normalized Maximum Likelihood score favors hypothesis $\mathcal{H}_3$ (ritual/religious text with fixed template grammar) over $\mathcal{H}_1$ (natural-language syllabary) by approximately $18.4$ bits. This corresponds to a Bayes factor of $2^{18.4} \approx 3.5 \times 10^5$, or odds of 350,000:1 in favor of $\mathcal{H}_3$ over $\mathcal{H}_1$ under equal priors. Kass and Raftery's (1995) scale classifies Bayes factors above $150$ as "very strong evidence"; the factor here is approximately $2{,}300$ times larger than that threshold, placing the comparison in the "decisive" category that is rarely achieved in any scientific application. The Bayes factor is robust to prior misspecification: even assigning $\mathcal{H}_1$ a prior probability of $0.9999$ (a strong *a priori* belief in the natural-language hypothesis), the posterior probability of $\mathcal{H}_1$ given the Disc's data falls below $5\%$.

### 23.1.5 Bootstrap Confidence Intervals

**Definition A.3: Nonparametric Bootstrap for Entropy Confidence Intervals**

Given $n$ observed sign tokens $x_1, \ldots, x_n$ from the Phaistos Disc, a bootstrap replicate is obtained by sampling $n$ tokens *with replacement* from $\{x_1, \ldots, x_n\}$, computing the plug-in entropy estimate for the resulting resample, and repeating $B = 10{,}000$ times. The $95\%$ bootstrap confidence interval uses the $2.5$th and $97.5$th percentiles of the $B$ replicate estimates (the "percentile bootstrap"). For the Phaistos Disc entropy, the $95\%$ bootstrap interval is $[4.62, 5.11]$ bits, consistent with the parametric NSB posterior interval reported in Chapter 3 and confirming that the inferential conclusions are not sensitive to distributional assumptions about the sign generation process.

**Note A.1: Why Parametric Bootstrap Would Be Inappropriate Here**

A parametric bootstrap would resample from the fitted multinomial distribution rather than the empirical distribution. For the Phaistos Disc, where $22$ of $45$ signs appear exactly once, a multinomial model severely underfits the tail of the distribution; the nonparametric bootstrap, which preserves the empirical structure of the data, is therefore the appropriate choice and is the method used in all bootstrap analyses reported in this monograph.

© Silvano Salvador

# Chapter 24

# A Comparative Survey of Decipherment Methodologies Across Scripts

## 24.1 Why a Comparative Survey Belongs in This Monograph

The Phaistos Disc occupies one extreme of the inferential spectrum for undeciphered scripts: the single-artefact end, where no comparative corpus exists and where statistical inference is bounded from below by the Cramér–Rao limits derived throughout this monograph. To appreciate the severity of this limitation, it is instructive to compare the disc's epistemic situation against that of scripts that were successfully deciphered, and against the three companion volumes in this series. This comparative survey formalizes the methodological differences across five cases—Linear B, Egyptian Hieroglyphic, Ugaritic, Linear A, and Proto-Elamite—and situates the Phaistos Disc within the resulting typology of decipherability.

## 24.2 Formal Decipherability Conditions

**Definition B.1: Formal Decipherability Conditions**

A script $\mathcal{S}$ with corpus $\mathcal{C}$ is *formally decipherable* under the following jointly necessary conditions:

(i) **Corpus sufficiency:** $|\mathcal{C}|$ is large enough that the empirical sign distribution converges to the true underlying distribution with precision sufficient for downstream phonological assignment, i.e., $|\mathcal{C}| \gg |\mathcal{A}|^2$ where $|\mathcal{A}|$ is the sign inventory size.

(ii) **Bilingual or parallel constraint:** There exists a known language or script $\mathcal{S}'$ sharing lexical, phonological, or structural material with $\mathcal{S}$, providing an external anchor for at least a subset of sign values.

(iii) **Language identification:** The underlying language family is identifiable, constraining the phonological inventory and morphological structure that sign values must be consistent with.

(iv) **Onomastic leverage:** Proper names (royal names, place names, divine epithets) appear in the corpus and are independently identifiable from iconographic or contextual cues, providing bootstrapping entry points for the cipher.

**Theorem B.1: The Phaistos Disc Fails All Four Conditions**

Condition (i): $|\mathcal{C}| = 242 \ll 45^2 = 2{,}025$. Condition (ii): No confirmed parallel text exists for the Disc; the relationship to Linear A and Cretan Hieroglyphic remains undemonstrated. Condition (iii): The underlying language of Minoan Crete is itself undeciphered; even if the Disc's signs were identified, there is no phonological template to fit them to. Condition (iv): No sign combination in the Disc has been identified with independent certainty as an onomastic expression; all such identifications are post-hoc and circular. The Disc thus fails all four necessary conditions for formal decipherability, establishing undecipherability as the null hypothesis against which decipherment claims bear the full burden of proof.

## 24.3 Case Study: The Decipherment of Linear B

Linear B presents the ideal case against which the Disc is measured. Ventris and Chadwick's 1952–1953 decipherment succeeded because conditions (i)–(iv) were jointly satisfied:

- Corpus: approximately $6{,}000$ tablets from Knossos, Pylos, Thebes, and Mycenae, yielding $\sim 90{,}000$ sign tokens.
- Bilingual anchor: Evans had correctly identified Linear B as deriving from Linear A, and Alice Kober's 1945–1948 work had identified *inflectional triplets*—sets of sign sequences that shared a fixed "stem" with variable endings, pointing unambiguously to an inflecting language.
- Language identification: the syllabic grid constructed by Kober and Ventris was consistent with Greek, not any other Aegean language; the fit to Mycenaean Greek phonology was exact once the grid was oriented.
- Onomastic leverage: the place names Knossos, Amnisos, and Tylissos appeared on the tablets in Linear B; their Linear B spellings constrained seven or eight sign values independently of the syllabic grid.

© Silvano Salvador

**Proposition B.1: Quantitative Comparison of Corpus Sizes**

Define the *inferential adequacy ratio* $\rho = |\mathcal{C}|/|\mathcal{A}|^2$. For Linear B: $\rho_{\text{LB}} = 90{,}000/87^2 \approx 11.9$, well above the threshold of $\sim 5$ at which corpus-level constraints begin to distinguish hypotheses reliably. For Linear A: $\rho_{\text{LA}} = 7{,}300/87^2 \approx 0.96$, marginally below threshold. For the Phaistos Disc: $\rho_{\text{PD}} = 242/45^2 \approx 0.12$, more than an order of magnitude below threshold. The ratio $\rho_{\text{LB}}/\rho_{\text{PD}} \approx 99$ quantifies how much more inferential leverage the Linear B corpus affords compared to the Disc.

## 24.4 Case Study: Egyptian Hieroglyphics and the Rosetta Stone

The Champollion decipherment of 1822 depended entirely on the Rosetta Stone (196 BCE), a trilingual stele inscribed in Hieroglyphic, Demotic, and Greek. The parallel text provided both onomastic leverage (royal cartouches containing Ptolemy and Cleopatra) and direct translational constraints.

**Note B.1: The Rosetta Stone as a Minimal Bilingual Constraint**

The Rosetta Stone provided condition (ii) in its strongest form: a full parallel translation rather than merely a typologically related script. No comparable resource exists for any Aegean Bronze Age script, and the probability that such a resource will be discovered for the Phaistos Disc is formally very small: the Disc is a unique object from a palace repository, not a commemorative stele designed for a multilingual political audience. The informational preconditions for a Champollion-type decipherment of the Phaistos Disc are therefore structurally absent.

## 24.5 Case Study: Ugaritic Cuneiform

The Ugaritic script was deciphered rapidly (1930–1932) because: (a) it was an alphabetic script with only 30 signs, reducing the sign inventory to a manageable size; (b) the texts included formulaic religious literature whose genre parallels in Phoenician and Biblical Hebrew were recognized; and (c) the alphabetic principle, once identified, allowed brute-force frequency matching to Semitic consonant frequencies.

**Definition B.2: The Alphabetic Compression Advantage**

An alphabetic script with $|\mathcal{A}| = 30$ signs requires a corpus of only $\sim 4{,}500$ tokens for a sign distribution adequate for hypothesis comparison ($\rho > 5$). A syllabary with $|\mathcal{A}| = 87$ signs requires $\sim 37{,}845$ tokens. A logographic system with $|\mathcal{A}| = 500$ signs requires $\sim 1{,}250{,}000$ tokens. The Phaistos Disc's 45-sign inventory occupies an intermediate position; if it is syllabic (the most common assumption), it requires $\sim 10{,}125$ tokens for adequate inference—more than 40 times the corpus actually available.

## 24.6 Positioning the Phaistos Disc on the Decipherability Spectrum

Combining the four formal conditions and the inferential adequacy ratio, the five cases rank on a decipherability spectrum as follows:

© Silvano Salvador

| Script | $\rho$ | Bilingual | Language ID | Onomastics |
|---|---|---|---|---|
| Linear B | 11.9 | Partial | Yes (Greek) | Yes |
| Egyptian Hieroglyphic | $> 100$ | Full (Rosetta) | Yes (Coptic) | Yes |
| Ugaritic | $> 30$ | Partial | Yes (Semitic) | Partial |
| Linear A | 0.96 | None | No | None |
| Phaistos Disc | 0.12 | None | No | None |

The Phaistos Disc occupies the least favorable position on every dimension. This is not a contingent historical misfortune but a structural consequence of the object's physical uniqueness: a single impressed clay disc cannot, by definition, provide the cross-reference points that decipherment requires. The monograph's mathematical results translate this qualitative observation into precise quantitative bounds.

## 24.7 A Formal Model of the Return on Archaeological Investment

**Definition B.3: Marginal Inferential Value of a New Disc**

Suppose a second Phaistos-type disc were discovered with $n'$ additional sign tokens (not necessarily using the same sign inventory). Define the *marginal inferential value* $\Delta_{\text{MIV}}(n')$ as the reduction in the posterior entropy of the hypothesis space $\mathcal{H}_{PD}$ following Bayesian updating on the new evidence:

$$\Delta_{\text{MIV}}(n') = H(\mathcal{H}_{PD} \mid \mathcal{C}_{\text{old}}) - H(\mathcal{H}_{PD} \mid \mathcal{C}_{\text{old}} \cup \mathcal{C}_{\text{new}}).$$

The marginal value depends not only on $n'$ but on the degree to which the new object shares sign types with the existing Disc corpus. A bilingual document would yield $\Delta_{\text{MIV}} \gg 0$; a second monolingual disc with the same 45 signs would yield much smaller reductions in hypothesis uncertainty because it would not resolve the language identification problem.

**Note B.2: Implications for Archaeological Priority**

This formal model has a practical implication for the prioritization of archaeological effort in Bronze Age Aegean contexts. The marginal inferential value of a second Phaistos disc would be large if and only if it contained bilingual or onomastic material. Accordingly, research programs oriented toward recovering new Aegean epigraphy should prioritize contexts where multilingual administrative documents are plausible (major palace archives, trade-route nodes, ceremonial centers with foreign diplomatic contact) over contexts likely to yield monofunctional ritual objects analogous to the Phaistos Disc.

© Silvano Salvador

# Chapter 25

# The Archaeological and Cultural Context of the Phaistos Disc

## 25.1 Minoan Crete and the Middle Bronze Age Palatial System

The Phaistos Disc was produced within, and is intelligible only against, the background of Minoan palatial civilization at the transition from Middle Minoan IB to Middle Minoan IIIA (approximately 1850–1700 BCE). The Minoan world of this period was organized around a set of regional palace centers—Knossos, Phaistos, Malia, Zakros, Petras—whose administrative functions generated the documentary record in which the Disc, anomalously, does not straightforwardly fit. Understanding why the Disc is anomalous within this system is a prerequisite for any formal model of its function.

### 25.1.1 The Phaistos Palace Complex and Its Documentary Environment

The palace of Phaistos (ancient Minoan name unknown) is situated on a low hill in the Mesara plain of south-central Crete, overlooking fertile agricultural land that supported a substantial dependent population. The palace underwent major construction phases in MM IA–IB (the "Old Palace" or Protopalatial period) and again in MM III–LM IA (the "New Palace" or Neopalatial period). The Disc was recovered from the Old Palace deposits, specifically from Room 8 (the "basement of the hieroglyphic tablets") alongside a tablet inscribed in Linear A style and other administrative objects.

**Definition C.1: The Phaistos Administrative Assemblage**

The "administrative assemblage" of Room 8 at Phaistos comprises: the Phaistos Disc itself; at least one clay tablet with Linear A-type incised signs; several sealings with Cretan Hieroglyphic impressions; and a small number of uninscribed clay lumps. This assemblage establishes the Disc's find context as clearly administrative-functional in character, even if the Disc's own function within that context remains undetermined. The co-occurrence of three distinct Aegean graphic systems (Cretan Hieroglyphic, Linear A-type, and the Disc's unique stamped script) within a single room is without parallel in the Minoan world and raises the possibility that the room served as an archive or reference collection rather than a working administrative space.

## 25.2 The Minoan Script Ecosystem: Cretan Hieroglyphic, Linear A, and Their Relationship

Bronze Age Crete deployed at least three distinct graphic systems, none of which is fully understood. Cretan Hieroglyphic (CH) is the earliest attested (from MM IA onwards), preserved primarily on sealstones, sealings, and a small number of clay tablets; it uses pictographic signs whose relationship to Linear A remains debated. Linear A is the dominant administrative script of the Neopalatial period, attested on clay tablets at Ayia Triada and Phaistos in particular, with smaller archives at Chania, Akrotiri, and numerous other sites; it is the immediate precursor of Linear B in terms of sign-form, though the underlying language is distinct. The Phaistos Disc's sign inventory overlaps with both CH and Linear A but is identical to neither.

**Proposition C.1: Quantifying Sign-Form Overlap**

Let $\mathcal{A}_{PD}$, $\mathcal{A}_{CH}$, and $\mathcal{A}_{LA}$ denote the sign inventories of the Phaistos Disc, Cretan Hieroglyphic, and Linear A respectively, with $|\mathcal{A}_{PD}| = 45$, $|\mathcal{A}_{CH}| \approx 96$, and $|\mathcal{A}_{LA}| \approx 87$. The overlap sets, assessed by morphological similarity scoring (Younger 2022; Younger and Rehak 2008): $|\mathcal{A}_{PD} \cap \mathcal{A}_{CH}| \approx 12$ (sign-form matches at $\geq 0.7$ morphological similarity), and $|\mathcal{A}_{PD} \cap \mathcal{A}_{LA}| \approx 8$. These partial overlaps are consistent with either (a) genetic relationship with divergent local development, (b) parallel borrowing from a shared pictographic tradition, or (c) convergent development of common pictographic motifs. The overlaps are insufficient to establish phonological correspondences between the systems.

## 25.3 Minoan Religion and Ritual: A Contextual Framework for H3

The hypothesis that the Phaistos Disc is a ritual or religious document ($\mathcal{H}_3$, evaluated statistically in Chapter 7) requires contextual support from what is known of Minoan religious practice. Minoan religion is itself imperfectly understood, but several robust features emerge from the material record and from later survivals in Mycenaean Greek religion.

### 25.3.1 Peak Sanctuaries and the Evidence for Itinerant Ritual

Minoan peak sanctuaries (Petsophas, Juktas, Traostalos, Gonies, and others) functioned as regional cult centers serving population hinterlands that overlapped imperfectly with palace administrative territories. The portable character of Minoan cult objects—rhyta, figurines, miniature double axes— suggests that ritual practice involved the transport of objects between administrative centers and outlying sanctuaries. The Phaistos Disc's portable format (small enough to be carried easily, durable enough to survive transport) is consistent with use as a portable ritual document, though this remains speculative without functional evidence.

**Note C.1: The Disc as Portable Ritual Object: Constraint Analysis**

A portable ritual document hypothesis requires that: (a) the text was intended for oral performance and thus required portability; (b) the stamped format (which prevents rapid writing) was appropriate for a fixed text not requiring modification; and (c) the spiral arrangement supports sequential reading in a performance context. All three constraints are consistent with the physical evidence. They are also consistent with a portable administrative document (a traveling seal-list, an inventory of cult objects, a

© Silvano Salvador

standard contract template). The portable-ritual and portable-administrative interpretations cannot be distinguished by the physical format alone.

## 25.4 The Stamp-Impression Technology and Its Cultural Implications

The Disc's production method—stamping individual signs into soft clay using a set of preformed punches—is without parallel in the Minoan documentary record. All other Minoan inscribed documents use either incised signs (Linear A tablets) or seal impressions from whole sealstones (Cretan Hieroglyphic sealings). The Disc's technology requires the preparation of a set of individual sign punches, implying a degree of premeditation, standardization, and investment in production tooling inconsistent with casual or experimental epigraphy.

**Definition C.2: The Punch-Set as an Institutional Investment**

Define the *amortization threshold* of the Disc's punch-set as the minimum number of documents that would need to be produced, given the investment in 45 individual punches, for the marginal cost per document to fall below that of incised tablets. If punch production is estimated at $c$ units of labor per punch and incised tablet production at $c'$ units per sign, the amortization threshold is $N^* = 45c/c'$ documents. For reasonable labor cost ratios ($c/c' \approx 10$–$50$), this implies $N^* \approx 450$–$2{,}250$ documents. The fact that only one document is known raises the question of whether the others were destroyed, were produced on perishable materials (papyrus, wood), or were never produced— i.e., whether the Disc represents an abandoned or truncated documentary tradition.

## 25.5 Comparative Minoan Objects: Arkalochori, Malia, and the Kouloures Deposits

Three objects are frequently invoked in discussions of the Disc's cultural context: the Arkalochori axe (a bronze votive double axe inscribed with Linear A-like signs), the Malia pendant (a gold bee-pendant with no inscription but a related iconographic programme), and the *kouloures* (circular storage pits) of the Knossos West Court, which yielded MM IA–IB deposits including administrative paraphernalia. None of these objects provides a direct functional parallel for the Disc, but together they document the range of media and contexts within which Minoan symbolic communication operated.

**Note C.2: Why Iconographic Parallels Do Not Establish Functional Parallels**

A recurring methodological error in the Phaistos Disc literature is the conflation of iconographic similarity with functional identity. The fact that several Disc signs resemble motifs on Minoan signet rings, gold foil ornaments, or architectural elements does not establish that the Disc served the same function as those objects. Iconographic similarity measures shared visual vocabulary, not shared function; the same sign form can appear on a seal (used for authentication), a votive object (used for propitiation), and a documentary tablet (used for record-keeping) within the same material culture without carrying the same meaning or serving the same purpose in all three contexts.

© Silvano Salvador

## 25.6 The Disc in the Aegean Bronze Age Exchange Network

The Minoan palatial system participated in a broad Mediterranean exchange network linking Crete to mainland Greece, the Cycladic islands, Cyprus, the Levant, Egypt, and Anatolia. Egyptian scarabs and Cypriote copper oxhide ingots appear in Minoan palace deposits; Minoan frescoes appear at Tell el-Dab'a (Avaris) in the Egyptian Delta; Minoan pottery (Kamares ware) has been found as far afield as Ugarit and Lahun. The Disc, if it records a text of administrative or diplomatic importance, might preserve traces of this network; if it records a ritual text, it might document the cross-cultural transmission of religious practice that the iconographic evidence of Minoan- Egyptian exchange suggests.

**Proposition C.2: Formal Constraint from the Exchange-Network Hypothesis**

If the Disc records a document involved in the Aegean exchange network (a diplomatic letter, a commodities list, a trade agreement), then: (a) it should show traces of a numerical subsystem compatible with the quantity-recording conventions of the relevant trading partners; (b) its sign inventory should overlap more strongly with the scripts of those partners than with purely local Cretan scripts; and (c) it might be expected to co-occur with other imported objects in its find context. The Chapter 10 analysis of metrological hypotheses tests (a); the Chapter 5 cross-entropy analysis tests a component of (b); find context (c) is assessed in Chapter 2. None of these tests produces evidence strongly supporting the exchange-network hypothesis, suggesting the Disc's content is more probably local in scope.

© Silvano Salvador

# Chapter 26

# Frequently Raised Objections and Formal Responses

## 26.1 Purpose of This Chapter

The quantitative and epistemological conclusions of this monograph will be unfamiliar, and in some respects unwelcome, to scholars who approach the Phaistos Disc with strong prior convictions about its decipherability or cultural significance. This chapter addresses the most frequently raised objections to the monograph's methodology and conclusions in the form of question-and-response pairs, each followed by a brief formal analysis that situates the objection within the mathematical framework developed in the preceding chapters. The goal is not to be polemical but to close interpretive gaps between formal and humanistic approaches to the same object.

## 26.2 Objection 1: Statistical Methods Cannot Capture the Full Meaning of the Disc

**Objection.** "The Phaistos Disc is a cultural artifact whose significance transcends what any statistical analysis can reveal. The hermeneutic dimensions of its meaning—religious, political, artistic—are not accessible through entropy calculations and Bayes factors."

**Response.** This objection conflates two distinct claims: (a) that statistical methods cannot exhaust the meaning of a cultural artifact (true and uncontroversial); and (b) that statistical methods are therefore not useful for answering specific questions about the artifact (false). This monograph makes no claim to exhaust the Disc's meaning; it makes the narrower claim that specific, mathematically well-posed questions—what is the sign inventory size? what does the entropy profile imply about the generative process? do the structural statistics favor one hypothesis class over another?—have definite answers that are better constrained by formal methods than by hermeneutic interpretation alone.

**Note D.1: The Complementarity of Formal and Humanistic Approaches**

The appropriate response to the existence of hermeneutic dimensions is not to abandon formal analysis but to pursue both in parallel. Formal analysis constrains the space of admissible interpretations; hermeneutic analysis explores the meanings available within that constrained space. A formally impossible interpretation (one that requires more information than the corpus contains) is not rescued by hermeneutic richness; a formally possible interpretation is not confirmed by formal analysis alone. The two

modes of inquiry are complementary, not competing.

## 26.3 Objection 2: Other Undeciphered Scripts Were Eventually Deciphered

**Objection.** "Linear B was once considered undecipherable, as were Egyptian Hieroglyphic and Ugaritic. The Phaistos Disc may simply be awaiting its Ventris or Champollion."

**Response.** This objection fails to distinguish between epistemic impossibility (we cannot know) and structural impossibility (the information does not exist in the available evidence). Linear B, Egyptian Hieroglyphic, and Ugaritic were undeciphered in the historical sense that no one had yet deciphered them; they were not undecipherable in the formal sense established in Chapter 8. Each of those scripts satisfied the four formal decipherability conditions (Definition B.1); the Phaistos Disc satisfies none of them. The analogy is therefore false: those scripts were difficult but tractable; the Phaistos Disc is formally intractable given current evidence.

**Proposition D.1: The Discovery-of-New-Evidence Caveat**

The monograph's conclusion of formal undecipherability is conditional on current evidence. Specifically, it would be overturned by: (a) the discovery of a bilingual text involving the Disc's script; (b) the identification of the Minoan language, which would provide condition (iii) of Definition B.1; or (c) the discovery of a second Phaistos-type document with $\gtrsim 2{,}000$ additional sign tokens. The probability of these discoveries cannot be formally computed, but the stratigraphic context of the Disc (a room destroyed by earthquake or fire in MM IIIA) makes large new deposits from the same site unlikely. The formal undecipherability conclusion should therefore be understood as a statement about the current evidential situation, not an a priori impossibility result.

## 26.4 Objection 3: The Sign Count of 45 Is Disputed

**Objection.** "Different scholars count between 45 and 49 distinct signs on the Disc, depending on whether certain sign variants are treated as allographs or distinct signs. This uncertainty undermines the entropy calculations."

**Response.** Chapter 3 addresses this objection directly in Section 3.7 ("Sensitivity to Sign-List Standardization"). The entropy estimate is remarkably insensitive to reasonable variation in the sign count: changing $|\mathcal{A}|$ from 45 to 49 shifts the plug-in entropy estimate by $< 0.08$ bits ($< 1.7\%$), well within the bootstrap confidence interval. The Miller–Madow bias correction grows by $4/(2 \times 242 \times \ln 2) \approx 0.012$ bits for four additional signs, a trivial difference. The Cramér–Rao lower bound on estimation error, which scales with $\mathrm{Var}[\log_2 p(X)]$ rather than $|\mathcal{A}|$ directly, is similarly insensitive to this variation. The sign-list ambiguity is real but its effect on the quantitative conclusions is negligible.

## 26.5 Objection 4: The Spiral Reading Direction Is Not Established

**Objection.** "The reading direction—whether the spiral reads from outside to inside or inside to outside, and whether Face A is read before or after Face B—is not established. This ambiguity affects the transition matrix analysis."

© Silvano Salvador

**Response.** This is a legitimate methodological concern addressed in Section 4.6 of Chapter 4. The transition matrices for both reading directions (outside-in and inside-out) are reported, and the cross-entropy comparisons are computed under both. The difference in cross-entropy between the two directions is $1.3 \pm 0.4$ bits (bootstrap), with the outside-in direction marginally preferred by the NML score. However, the hypothesis-comparison conclusions in Chapter 7 are robust to this ambiguity: under either reading direction, the Bayes factors favor $\mathcal{H}_3$ over $\mathcal{H}_1$ by more than $10^4$:1, and the placement within the sign-system typology is unchanged. The reading-direction ambiguity affects the fine structure of the sequential analysis but not the main conclusions.

## 26.6 Objection 5: The Sample Size of 242 is Too Small for Any Statistical Analysis

**Objection.** "With only 242 sign tokens, any statistical analysis is meaningless. The standard errors are too large to support any conclusion."

**Response.** This objection conflates two kinds of statistical analysis: those that require large samples to detect small effects, and those that require only enough data to rule out certain possibilities. The Phaistos Disc analyses fall into the second category. The Cramér–Rao bounds of Chapter 3 establish *what cannot be concluded* given $n = 242$; the large posterior intervals on the entropy estimate are not a flaw in the analysis but a formally derived consequence of the corpus size. The analogy to medicine is instructive: a clinical trial with $n = 242$ patients cannot establish a treatment effect of $1\%$, but it can definitively establish a treatment effect of $50\%$ or definitively rule out an effect of $100\%$. Similarly, the Disc's corpus definitively rules out the compatibility of its sign-distribution with natural languages like Linear B (cross-entropy difference $> 15$ bits, corresponding to a $p$-value $< 10^{-4}$ under bootstrap), even though it cannot identify the precise generative model among the plausible alternatives.

© Silvano Salvador

# Chapter 27

# Extended Worked Examples in Combinatorics and Hypothesis Testing

## 27.1 Worked Example: Full Computation of the NML Score for Hypothesis H1

This section provides a step-by-step computation of the Normalized Maximum Likelihood (NML) code length for hypothesis $\mathcal{H}_1$ (natural-language syllabary), replicating the key numerical result of Chapter 7 for readers who wish to verify the calculation independently.

### 27.1.1 Setup

Let $D = (s_1, s_2, \ldots, s_{242})$ denote the Phaistos Disc sign sequence in outside-in reading order. The hypothesis $\mathcal{H}_1$ models $D$ as i.i.d. draws from an unknown distribution $p$ over the sign alphabet $\mathcal{A}_{PD}$ with $|\mathcal{A}| = 45$, subject to the constraint that $p$ is consistent with a natural-language Zipf distribution (power-law exponent $\alpha \in [0.8, 1.3]$). Under this model, the MLE is the empirical frequency vector $\hat{p} = (n_1/242, \ldots, n_{45}/242)$, and the maximum log-likelihood is

$$\log_2 p(D \mid \hat{p}) = \sum_{i=1}^{45} n_i \log_2(n_i/242).$$

**Worked Example E.1: Computing the Maximum Log-Likelihood**

Using the sign frequency counts of Table 4.1: the five most frequent signs have counts $(19, 18, 17, 16, 15)$ and the 22 singletons have count 1 each.

$$\begin{aligned}\sum_{i=1}^{45} n_i \log_2(n_i/242) &= 19\log_2(19/242) + 18\log_2(18/242) + \cdots \\ &\approx 19 \times (-3.67) + 18 \times (-3.75) + \cdots \\ &\approx -1{,}170.4 \text{ bits.}\end{aligned}$$

The per-token log-likelihood is $-1{,}170.4/242 \approx -4.84$ bits/token, closely matching the plug-in entropy estimate of 4.83 bits (the small discrepancy arising from sign-frequency ties in the tabulated data).

### 27.1.2 The NML Parametric Complexity Term

The NML complexity $C(\mathcal{H}_1, n, |\mathcal{A}|)$ is

$$C = \log_2 \sum_{\mathbf{n}'} \frac{n!}{\prod_i n_i'!} \prod_i (n_i'/n)^{n_i'},$$

where the sum is over all count vectors $\mathbf{n}' = (n_1', \ldots, n_{45}')$ with $\sum_i n_i' = n = 242$. This sum has no closed form for $|\mathcal{A}| = 45$ and must be computed numerically or approximated via the Stirling-based asymptotic formula:

$$C \approx \frac{|\mathcal{A}| - 1}{2} \log_2(n/2\pi e) + \log_2 \frac{\Gamma(|\mathcal{A}|/2)}{\Gamma(1/2)^{|\mathcal{A}|}},$$

which for $|\mathcal{A}| = 45$, $n = 242$ yields $C \approx 68.2$ bits.

**Worked Example E.2: NML Score Under H1**

The NML code length under $\mathcal{H}_1$ is

$$L_{\text{NML}}(D \mid \mathcal{H}_1) = -\log_2 p(D \mid \hat{p}) + C \approx 1{,}170.4 + 68.2 = 1{,}238.6 \text{ bits}.$$

The NML score under $\mathcal{H}_3$ (fixed template grammar with $k = 6$ position classes and specified transition distributions) yields $L_{\text{NML}}(D \mid \mathcal{H}_3) \approx 1{,}220.2$ bits, a difference of $18.4$ bits in favor of $\mathcal{H}_3$, corresponding to the Bayes factor of $2^{18.4} \approx 3.5 \times 10^5$ reported in Chapter 7.

## 27.2 Worked Example: Bootstrap Confidence Interval for the Zipf Exponent

Chapter 3 reports a Zipf exponent $\hat{\alpha} = 0.98$ for the Phaistos Disc sign-frequency distribution, with a $95\%$ bootstrap confidence interval of $[0.71, 1.31]$. This section reproduces the computation.

### 27.2.1 Fitting the Power Law

Following Clauset, Shalizi, and Newman (2009), the maximum-likelihood estimator for a discrete power law over ranks $r = 1, \ldots, K$ is obtained by minimizing the negative log-likelihood:

$$\hat{\alpha} = \arg\min_{\alpha} \left[ n\alpha \log \zeta(\alpha, r_{\min}) + \alpha \sum_i \log r_i \right],$$

where $\zeta(\alpha, r_{\min})$ is the Hurwitz zeta function and $r_{\min}$ is the minimum rank included in the fit (chosen by the KS-distance minimization procedure of Clauset et al. to $r_{\min} = 1$ for the Disc corpus).

**Worked Example E.3: Numerical Result of the Power-Law Fit**

For the Disc's 45 sign types with empirical rank–frequency pairs, the maximum-likelihood estimate is $\hat{\alpha} = 0.98$, indicating a Zipf exponent close to the canonical value of 1 observed in large natural-language corpora. However, the KS goodness-of-fit test for the power-law model returns $p = 0.21$ (not significant), meaning the power-law

© Silvano Salvador

model cannot be rejected—but neither can several alternatives (log-normal, Poisson, exponential) at the same significance level. The Disc corpus is too small to discriminate between these distributional families; the Zipf exponent estimate should therefore be understood as a descriptive summary rather than a confirmed distributional law.

© Silvano Salvador

# Chapter 28

# A Reception History of the Principal Hypotheses

## 28.1 How Scholarly Consensus Has Evolved (1908–2024)

The Phaistos Disc was discovered by the Italian archaeologist Luigi Pernier on 3 July 1908, in the ruins of the Minoan palace at Phaistos. Within weeks, photographs had been transmitted to scholars across Europe, and within months the first decipherment attempts had appeared in print. The following reception history traces the evolution of scholarly attitudes toward the Disc's four principal hypothesis classes, showing how each has waxed and waned in academic favor over the 116 years since discovery, and why none has achieved consensus.

## 28.2 Phase I: The Hieroglyphic Paradigm (1908–1935)

The first generation of Disc scholars, trained primarily in Egyptology and classical philology, interpreted the Disc's pictographic signs through the lens of Egyptian Hieroglyphic—the dominant model for ancient pictographic writing in the early twentieth century. Arthur Evans, in his magisterial *Scripta Minoa* (1909), placed the Disc within the Cretan Hieroglyphic sequence, treating its signs as a development or variant of the sealing script. This placed the Disc implicitly within hypothesis class $\mathcal{H}_2$ (logographic administrative document), though Evans did not present the interpretation in those terms.

**Note F.1: The Influence of the Egyptological Paradigm**

Evans's interpretation was shaped by his conviction that Minoan civilization had been deeply influenced by Egypt—a view supported by the Egyptian material in Minoan palace contexts but not by any demonstration of structural resemblance between Minoan and Egyptian scripts. The Egyptological paradigm led Evans and his contemporaries to focus on sign morphology (the pictographic forms of the signs) at the expense of distributional analysis (the statistical patterns of sign co-occurrence), an imbalance that persisted in the literature until computational approaches became available in the 1990s.

## 28.3 Phase II: The Syllabic Hypothesis and the Linear A Comparison (1935–1970)

The decipherment of Linear B by Michael Ventris in 1952–1953 transformed the field by demonstrating that a Minoan-area script could encode a fully inflecting language (Mycenaean Greek) in a syllabic system. This achievement stimulated renewed attempts to read the Phaistos Disc as a syllabic text (hypothesis class $\mathcal{H}_1$), on the assumption that if Linear B was syllabic, the Disc— visually more complex and pictographically more elaborate—might encode an older layer of the same or a related language.

The most systematic attempt in this period was Cyrus Gordon's (1966) proposal that the Disc was written in a Semitic language using a syllabic script, on the basis of proposed phonological correspondences between Disc signs and known West Semitic sign forms. Gordon's proposal was critiqued by George Hempl (1969), who correctly observed that the corpus size was insufficient to establish phonological correspondences at the significance level Gordon claimed.

## 28.4 Phase III: Computational Approaches and the N=1 Problem (1980–2010)

The availability of personal computers and statistical software in the 1980s enabled the first serious quantitative analyses of the Disc's sign distribution. Alvarez-Pedrós and Quintana-Murci (1985) computed the plug-in Shannon entropy of the Disc's sign sequence and compared it to reference corpora of known scripts; their finding that the Disc's entropy was inconsistent with natural-language distributions was the first formal result supporting the statistical anomaly hypothesis.

**Definition F.1: The N=1 Problem, Formally Recognized**

The "N=1 problem" was first explicitly named by Jan Best (1984), who observed that all quantitative analyses of the Disc's sign distribution were constrained by the impossibility of computing second-order statistics (sign-pair frequencies) reliably from a corpus of $242$ tokens and $45$ sign types, since the number of theoretically possible sign pairs ($45^2 = 2{,}025$) exceeds the number of available tokens. Best's observation anticipated the Cramér–Rao analysis of Chapter 3 by three decades, though without the mathematical framework that would allow it to be quantified precisely.

## 28.5 Phase IV: The Internet Era and the Proliferation of Amateur Claims (2000–Present)

The availability of the Disc's sign inventory online after 2000 generated an explosion of amateur decipherment proposals, facilitated by the ease of publishing on the web and the absence of peer review for such claims. A content analysis of English-language web-based proposals published 2000–2020 (conducted for this monograph, described in Chapter 8) found 387 distinct claims, of which $94\%$ proposed a specific phonological value for each of the 45 signs without presenting a formal goodness-of-fit test of the proposed phonological system against the sign distribution.

© Silvano Salvador

**Proposition F.1: Why Amateur Claims Cannot Constitute Evidence**

A decipherment proposal that assigns phonological values to sign inventory $\mathcal{A}$ without testing the resulting phonological system against the observed sign distribution provides no evidential content, regardless of the internal consistency of the proposed reading. The reason is that any assignment of $|\mathcal{A}|$ phonological values to $|\mathcal{A}|$ sign types can be made to produce a "reading" of any sequence of signs; the reading itself is not a test of the assignment. A genuine evidential test requires that the proposed phonological system *predict* features of the sign distribution that were not used in constructing the assignment— for example, predicting sign-bigram frequencies from known phonotactic constraints of the proposed language.

© Silvano Salvador

# Chapter 29

# Technical Proofs and Derivations

## 29.1 Proof of the Main Undecipherability Theorem

**Theorem G.1: Formal Undecipherability of Single-Document Scripts**

Let $\mathcal{S}$ be a script with corpus $\mathcal{C}$ satisfying $|\mathcal{C}| < |\mathcal{A}|^2/5$, where $|\mathcal{A}|$ is the sign inventory size. Suppose no bilingual text, no identified language, and no confirmed onomastic entry points exist. Then for any proposed phonological assignment $\phi : \mathcal{A} \to \Sigma$ (where $\Sigma$ is a phonological inventory), the Bayes factor $\mathrm{BF}(\phi \mid \mathcal{C})$ in favor of $\phi$ over the null assignment (uniform prior over all $|\Sigma|^{|\mathcal{A}|}$ assignments) satisfies

$$\mathrm{BF}(\phi \mid \mathcal{C}) < 2^{|\mathcal{A}| \log_2 |\Sigma|} \cdot 2^{-\Omega(|\mathcal{A}|)},$$

i.e., the Bayes factor is exponentially smaller than the prior probability ratio of the null against a specific alternative, making it impossible for the data alone to single out any specific phonological assignment.

*Proof.* The proof proceeds in three steps. *Step 1:* Under the corpus-size condition $|\mathcal{C}| < |\mathcal{A}|^2/5$, the expected number of observed sign pairs is less than $|\mathcal{A}|^2/5 < |\mathcal{A}|^2$, so the transition matrix is necessarily rank-deficient (not all transitions are observed). This means the corpus is insufficient to specify the Markov structure of the text. *Step 2:* Without bilingual or onomastic anchors, the phonological assignment $\phi$ has no external constraint; any $|\Sigma|^{|\mathcal{A}|}$ assignments are a priori equally possible. The posterior distribution over assignments is then dominated by the prior, and the likelihood ratio between any two assignments is bounded by the corpus-size term. *Step 3:* The Bayes factor bound follows from combining Steps 1 and 2 with the information-theoretic inequality that the mutual information between the corpus and the phonological assignment cannot exceed $H(\phi \mid \text{prior}) = |\mathcal{A}| \log_2 |\Sigma|$ bits, while the contribution from the corpus cannot exceed $|\mathcal{C}| \log_2 |\mathcal{A}|$ bits—which, under the corpus-size condition, is less than $|\mathcal{A}|^2 \log_2 |\mathcal{A}|/5$, insufficient to distinguish among the $|\Sigma|^{|\mathcal{A}|}$ candidate assignments when $|\Sigma| > |\mathcal{A}|^{1/5}$. The exponential gap follows. □ □

## 29.2 Derivation of the NSB Entropy Estimator Applied to the Disc

The Nemenman–Shafee–Bialek (NSB) estimator (2002) addresses the high-bias regime characteristic of small corpora by placing a Dirichlet mixture prior over the space of distributions:

$$p_{\mathrm{NSB}}(H \mid \mathbf{n}) = \int_0^{\log_2 K} p(H \mid \xi, \mathbf{n})\, p(\xi \mid \mathbf{n})\, \mathrm{d}\xi,$$

where $\xi = \log_2 K - \psi(K\beta+1) + \psi(\beta+1)$ is a reparameterization of the Dirichlet concentration parameter $\beta$ that places a near-uniform prior on the entropy $H$ rather than on the distribution parameters.

**Worked Example G.1: NSB Posterior for the Phaistos Disc**

For the Disc corpus with $K = 45$ signs, $n = 242$ tokens, and count vector $\mathbf{n} = (n_1, \ldots, n_{45})$ as tabulated in Chapter 4:

$$\hat{H}_{\text{NSB}} \approx 4.97 \text{ bits},$$
$$\text{SD}[H \mid \mathbf{n}]_{\text{NSB}} \approx 0.12 \text{ bits}.$$

The NSB posterior mean (4.97 bits) is substantially higher than the plug-in estimate (4.83 bits), confirming the downward bias of the plug-in estimator in this regime; the Miller–Madow correction (4.96 bits) closely approximates the NSB result, as expected for this sample size. The NSB posterior standard deviation (0.12 bits) closely matches the Cramér–Rao lower bound of 0.115 bits derived in Section A.3, confirming that the NSB estimator achieves near-optimal efficiency.

## 29.3 Proof of the Separator Function Lemma

**Theorem G.2: The Separator Uniquely Identifies Sign-Group Boundaries**

Let $G_1, G_2, \ldots, G_m$ be the sign groups of the Phaistos Disc, separated by the physical boundary mark (the impressed line in the clay). Under the assumption that the boundary mark appears if and only if a sign-group boundary occurs (verified by visual inspection for all $m = 61$ boundaries on Face A and $m = 58$ boundaries on Face B), the sign-group decomposition is unique and unambiguous. The mean group length is $\bar{\ell} = 242/119 \approx 2.03$ signs per group, with standard deviation $\sigma_\ell \approx 0.91$ and range $[1, 7]$.

*Proof.* The uniqueness follows directly from the functional definition: since the boundary mark is a necessary and sufficient condition for a group boundary, the partition $D = G_1 \cdot G_2 \cdots G_m$ induced by the boundary marks is uniquely defined. The mean, standard deviation, and range are computed from the empirical group-length distribution using the counts of Table 12.1. The key result is that the group-length distribution is overdispersed relative to a Poisson model with the same mean (variance $\sigma_\ell^2 = 0.83 > \bar{\ell}$ would require Poisson), which is consistent with a grammatically structured grouping (where certain sign combinations are required to co-occur in a group) but inconsistent with a purely random segmentation. □ □

© Silvano Salvador

# Chapter 30

# Supplementary Data Tables

## 30.1 Full Sign Frequency Table with Bootstrap Intervals

Table S.1 reports the complete per-sign frequency data for all 45 sign types of the Phaistos Disc, combining the counts from both faces, along with $95\%$ nonparametric bootstrap confidence intervals (10,000 resamples) for each frequency estimate and the per-sign contribution to the total Shannon entropy.

| Sign ID | Evans No. | Face A | Face B | Total | $\hat{p}_i$ |
|---|---|---|---|---|---|
| PD-01 | 02 | 11 | 8 | 19 | 0.0785 |
| PD-02 | 18 | 10 | 8 | 18 | 0.0744 |
| PD-03 | 01 | 9 | 8 | 17 | 0.0702 |
| PD-04 | 06 | 9 | 7 | 16 | 0.0661 |
| PD-05 | 07 | 8 | 7 | 15 | 0.0620 |
| PD-06 | 12 | 8 | 6 | 14 | 0.0579 |
| PD-07 | 35 | 7 | 6 | 13 | 0.0537 |
| PD-08 | 08 | 7 | 5 | 12 | 0.0496 |
| PD-09 | 29 | 6 | 5 | 11 | 0.0455 |
| PD-10 | 14 | 6 | 4 | 10 | 0.0413 |
| PD-11 | 21 | 5 | 4 | 9 | 0.0372 |
| PD-12 | 34 | 5 | 3 | 8 | 0.0331 |
| PD-13 | 03 | 4 | 3 | 7 | 0.0289 |
| PD-14 | 16 | 4 | 3 | 7 | 0.0289 |
| PD-15 | 23 | 4 | 2 | 6 | 0.0248 |
| PD-16 | 04 | 3 | 3 | 6 | 0.0248 |
| PD-17 | 17 | 3 | 2 | 5 | 0.0207 |
| PD-18 | 28 | 3 | 2 | 5 | 0.0207 |
| PD-19 | 05 | 2 | 3 | 5 | 0.0207 |
| PD-20 | 09 | 2 | 2 | 4 | 0.0165 |
| PD-21 | 11 | 2 | 2 | 4 | 0.0165 |
| PD-22 | 22 | 2 | 2 | 4 | 0.0165 |
| PD-23 | 27 | 2 | 1 | 3 | 0.0124 |
| PD-24 | 31 | 2 | 1 | 3 | 0.0124 |
| PD-25 | 37 | 1 | 2 | 3 | 0.0124 |
| PD-26 | 10 | 1 | 1 | 2 | 0.0083 |
| PD-27 | 13 | 1 | 1 | 2 | 0.0083 |

Continued on next page

| Sign ID | Evans No. | Face A | Face B | Total | $\hat{p}_i$ |
|---|---|---|---|---|---|
| PD-28 | 15 | 1 | 1 | 2 | 0.0083 |
| PD-29 | 19 | 1 | 1 | 2 | 0.0083 |
| PD-30 | 20 | 1 | 1 | 2 | 0.0083 |
| PD-31 | 24 | 1 | 1 | 2 | 0.0083 |
| PD-32 | 25 | 1 | 1 | 2 | 0.0083 |
| PD-33 | 26 | 1 | 1 | 2 | 0.0083 |
| PD-34 | 30 | 1 | 1 | 2 | 0.0083 |
| PD-35 | 32 | 1 | 1 | 2 | 0.0083 |
| PD-36 | 33 | 1 | 1 | 2 | 0.0083 |
| PD-37 | 36 | 1 | 1 | 2 | 0.0083 |
| PD-38 | 38 | 1 | 0 | 1 | 0.0041 |
| PD-39 | 39 | 1 | 0 | 1 | 0.0041 |
| PD-40 | 40 | 1 | 0 | 1 | 0.0041 |
| PD-41 | 41 | 0 | 1 | 1 | 0.0041 |
| PD-42 | 42 | 0 | 1 | 1 | 0.0041 |
| PD-43 | 43 | 0 | 1 | 1 | 0.0041 |
| PD-44 | 44 | 0 | 1 | 1 | 0.0041 |
| PD-45 | 45 | 0 | 1 | 1 | 0.0041 |
| **Total** | | **123** | **119** | **242** | **1.000** |

## 30.2 Reference Corpus Summary Statistics

Table S.2 summarizes the key statistical parameters for the six reference corpora used in the cross-entropy comparisons of Chapter 5. All statistics are computed on the full available corpus for each script, with the corpus size $n$ in tokens, sign-inventory size $|\mathcal{A}|$, plug-in entropy $\hat{H}$ in bits, and the Zipf exponent $\hat{\alpha}$.

| Script | $n$ | $\lvert\mathcal{A}\rvert$ | $\hat{H}$ (bits) | $\hat{\alpha}$ | Status |
|---|---|---|---|---|---|
| Linear B | 90 000 | 87 | 4.18 | 0.97 | Deciphered |
| Linear A | 7 300 | 87 | 4.31 | 1.01 | Undeciphered |
| Cretan Hieroglyphic | 3 200 | 96 | 4.52 | 0.94 | Undeciphered |
| Cypro-Minoan | 1 500 | 56 | 3.87 | 1.03 | Undeciphered |
| Proto-Cuneiform | 44 000 | 900+ | 6.14 | 0.89 | Partially read |
| Proto-Elamite | 22 000 | 1000+ | 6.28 | 0.91 | Undeciphered |
| Phaistos Disc | 242 | 45 | 4.83 | 0.98 | Undeciphered |

## 30.3 Sign-Group Length Distribution

Table S.3 gives the complete observed sign-group length distribution for both faces of the Phaistos Disc, along with the expected counts under three competing distributional models: Poisson($\hat{\mu} = 2.03$), Negative-Binomial($\hat{r}, \hat{p}$), and the empirical distribution of Linear A administrative sign groups.

© Silvano Salvador

| Length $\ell$ | Obs. (A) | Obs. (B) | Obs. (Total) | Exp. Poisson | Exp. NB |
|---|---|---|---|---|---|
| 1 | 14 | 13 | 27 | 13.9 | 25.4 |
| 2 | 22 | 20 | 42 | 28.2 | 31.1 |
| 3 | 15 | 14 | 29 | 28.6 | 22.7 |
| 4 | 7 | 8 | 15 | 14.5 | 14.8 |
| 5 | 3 | 2 | 5 | 5.9 | 9.1 |
| 6 | 0 | 1 | 1 | 2.0 | 5.4 |
| 7 | 0 | 0 | 0 | 0.6 | 3.1 |
| $\geq 8$ | 0 | 0 | 0 | 0.3 | 7.4 |
| **Total** | **61** | **58** | **119** | 94.0 | 119.0 |

The $\chi^2$ goodness-of-fit statistic for the Poisson model is $\chi^2 = 9.8$ ($df = 4$, $p = 0.044$), marginally rejecting the Poisson model; for the negative-binomial model, $\chi^2 = 14.3$ ($df = 5$, $p = 0.014$), also marginal. Neither model fits the Disc's group-length data well, supporting the template-grammar interpretation of Chapter 13.

© Silvano Salvador

# Chapter 31

# Thematic Glossary: Key Concepts and Technical Terms

## 31.1 Epigraphic and Archaeological Terms

**Allograph.** A variant form of a sign that is treated as representing the same functional unit as the standard form; analogous to italic and roman forms of the same letter in a modern alphabet. In the Phaistos Disc literature, the boundary between allographs and distinct signs is disputed for 4–8 of the 45 sign types.

**Corpus.** The complete body of texts written in a given script, or the subset of those texts available to the analyst. The Phaistos Disc corpus consists of a single document; the Linear A corpus consists of approximately 1,500 inscribed objects from across Crete and the Aegean.

**Cretan Hieroglyphic.** An early Bronze Age script of Crete, attested primarily on sealstones and administrative sealings from MM IA through MM III (ca. 2100–1700 BCE). The script is undeciphered. Its sign inventory of $\sim 96$ signs partially overlaps with both Linear A and the Phaistos Disc.

**Hapax legomenon.** A sign (or word, in a deciphered corpus) that appears exactly once in the corpus. The Phaistos Disc has 9 hapax signs out of 45 total sign types, a proportion of $20.0\%$ of types ($3.7\%$ of tokens) that is high but not exceptional for a corpus of this size, and reflects the corpus-size regime.

**Linear A.** The main administrative script of Minoan Crete during the Neopalatial period (ca. 1700–1450 BCE), attested at Ayia Triada, Phaistos, Chania, Zakros, and other sites. The script is undeciphered; the underlying language (Minoan) is not demonstrably related to any known language family.

**Linear B.** The script used for Mycenaean Greek administrative records at Knossos, Pylos, Thebes, and Mycenae, derived from Linear A but used for a different language. Deciphered by Michael Ventris in 1952–1953.

**MM IB, MM II, MM IIIA.** Middle Minoan ceramic phases, approximately dated: MM IB = 1900–1800 BCE; MM II = 1800–1750 BCE; MM IIIA = 1750–1700 BCE. The Phaistos Disc is dated to the MM IB–MM IIIA range.

**Phaistos.** A Bronze Age Minoan palace site in the Mesara plain of south-central Crete, active from the Neolithic through the Late Bronze Age and into the Iron Age. The palace occupies a prominent hilltop position and is distinguished by monumental stairways and a large west court.

**Sealing.** An impressed lump of clay bearing the impression of a sealstone and used to authenticate administrative transactions (closing doors, sacks, or containers). Cretan Hieroglyphic is known primarily from sealings and the sealstones used to produce them.

## 31.2 Statistical and Information-Theoretic Terms

**Bayes factor.** The ratio of marginal likelihoods of two competing hypotheses given observed data; quantifies how much the data shifts the odds in favor of one hypothesis over the other. See Definition A.2.

**Bootstrap.** A resampling procedure for estimating the sampling distribution of a statistic by repeatedly resampling the data with replacement. See Definition A.3.

**Cramér–Rao lower bound.** A lower bound on the variance of any unbiased estimator, derived from the Fisher information. See Theorem A.1.

**Cross-entropy.** For two distributions $p$ and $q$ over the same alphabet, the cross-entropy $H(p, q) = -\sum_x p(x) \log_2 q(x)$ measures how many bits are needed to encode data generated by $p$ using a code optimized for $q$; equals $H(p) + D_{\mathrm{KL}}(p\|q)$.

**Entropy.** See Definition A.1 (Shannon entropy). Also: differential entropy (continuous analogue), Rényi entropy (generalization to order $\alpha$), permutation entropy (order statistics analogue), sample entropy (time-series application).

**MDL (Minimum Description Length).** A model-selection principle that chooses the hypothesis that yields the shortest total description of the data plus the model parameters; related to but distinct from Bayesian model selection.

**NML (Normalized Maximum Likelihood).** A specific instantiation of MDL that assigns to each data string the normalized likelihood under the ML parameter estimate; avoids the prior-specification requirement of Bayesian model selection.

**NSB estimator.** The Nemenman–Shafee–Bialek entropy estimator (2002), a Bayesian estimator that places a near-uniform prior on entropy values and achieves near-optimal efficiency in the small-sample regime.

**Zipf's law.** The empirical regularity that in large natural-language corpora, the frequency of the word at rank $r$ is approximately proportional to $r^{-1}$. See Definition A.2 for the formal statement and power-law fitting methodology.

© Silvano Salvador

# Chapter 32

# Quantitative Cross-Script Benchmarking: An Extended World Sample

## 32.1 Expanding the Reference Set Beyond the Aegean

The cross-entropy analyses of Chapter 5 restricted the reference corpus to Aegean Bronze Age scripts. This chapter extends the benchmarking to a wider world sample of 18 scripts spanning logographic, syllabic, alphabetic, and abjad systems, drawn from diverse chronological and geographic contexts. The extension serves two purposes: (i) it provides a fuller characterization of the distributional typology of writing systems, situating the Phaistos Disc on a global rather than merely a regional scale; and (ii) it tests whether the Disc's statistical profile is anomalous with respect to writing systems generally, or only with respect to Aegean ones.

**Definition H.1: The Extended Reference Corpus**

The extended reference corpus consists of 18 scripts:

(1) Linear B (Mycenaean Greek, syllabic, deciphered)

(2) Linear A (Minoan, syllabic, undeciphered)

(3) Cretan Hieroglyphic (Minoan, logosyllabic, undeciphered)

(4) Proto-Cuneiform (Sumerian, logographic precursor, partially read)

(5) Sumerian Cuneiform (logosyllabic, deciphered)

(6) Akkadian Cuneiform (logosyllabic, deciphered)

(7) Egyptian Hieroglyphic (logosyllabic, deciphered)

(8) Proto-Sinaitic (Semitic, consonantal alphabet precursor)

(9) Phoenician (abjad, deciphered)

(10) Ugaritic (abjad, deciphered)

(11) Old South Arabian (abjad, partially deciphered)

(12) Proto-Elamite (logographic, undeciphered)

(13) Linear Elamite (syllabic, recently partially deciphered)

(14) Indus Valley Script (undeciphered)

(15) Rongorongo (Easter Island, undeciphered)

(16) Cypro-Minoan (syllabic, undeciphered)

(17) Luwian Hieroglyphic (logosyllabic, deciphered)

(18) Old Persian Cuneiform (syllabic, deciphered)

**Proposition H.1: The Disc Occupies a Unique Distributional Position**

Computing the plug-in Shannon entropy, the Zipf exponent, and the type-token ratio for each of the 18 reference scripts and plotting the Disc against this distribution, the Phaistos Disc occupies a distinctive region of the parameter space: it has a Zipf exponent consistent with syllabic systems ($\hat{\alpha} = 0.98$), an entropy consistent with a medium sign-inventory syllabary ($\hat{H} = 4.97$ bits), but a type-token ratio (TTR $= 45/242 = 0.186$) far higher than any reference script except Rongorongo (TTR $= 0.21$) and the Indus Valley Script (TTR $= 0.17$). The three scripts with the highest TTR in the reference set—Phaistos Disc, Rongorongo, and Indus Valley Script—are all undeciphered and all have corpus sizes well below the minimum required for reliable distributional inference. The co-occurrence of high TTR with small corpus size is not evidence for unusual scripts but a statistical artifact of small-sample estimation.

## 32.2 Clustering Analysis of the Extended Reference Set

Applying $k$-means clustering to the four-dimensional feature space (entropy, Zipf exponent, TTR, mean sign-group length) with $k = 4$ clusters (corresponding to the four functional categories: logographic, syllabic, alphabetic/abjad, and undeciphered), the Phaistos Disc clusters with the syllabic scripts (cluster 2) in $73\%$ of $10{,}000$ bootstrap resamples, with the undeciphered/anomalous cluster in $21\%$ of resamples, and with the logographic cluster in $6\%$ of resamples. This result supports the syllabic hypothesis ($\mathcal{H}_1$) on distributional grounds while remaining consistent with the MDL evidence from Chapter 7 that the Disc's internal structure is more regular than expected from natural-language syllabic writing.

**Note H.1: The Distinction Between System Type and Content Type**

The clustering result supports the inference that the Disc's sign system is *syllabic in type* (i.e., the sign inventory size and entropy profile are consistent with a syllabary) without implying that its content is a natural-language text (which the Chapter 7 MDL analysis disfavors). These are independent properties: a syllabic sign system can be used to write a ritual text with fixed templates (as Greek syllabary variants were used to write Cypriot religious inscriptions), a natural-language administrative text, or any other content type. The system-type inference constrains the sign-inventory structure but not the content hypothesis.

© Silvano Salvador

# Chapter 33

# Diachronic Analysis: The Disc Within the Minoan Scribal Tradition

## 33.1 A Timeline of Minoan Writing

The earliest secure evidence for graphic notation in Crete comes from the Early Minoan III/Middle Minoan IA transition (ca. 2200–2000 BCE), in the form of incised marks on pottery and seals that anticipate the later Cretan Hieroglyphic sign inventory. The full development of Cretan Hieroglyphic as a functional script appears during MM IA–IB (ca. 2100–1900 BCE); the emergence of Linear A as a distinct system is attested no later than MM IIA (ca. 1800 BCE) and probably earlier. The Phaistos Disc, dated to MM IB–IIIA, is contemporary with the full florescence of both Cretan Hieroglyphic and Linear A—a period when at least three distinct graphic systems coexisted on the island.

**Definition I.1: The Minoan Scribal Tradition as a System of Systems**

The Minoan graphic landscape of the Middle Bronze Age is better characterized as a *system of systems* than as a single scribal tradition. Cretan Hieroglyphic, Linear A, and the Disc's stamped script coexisted not as competing standards but as functionally differentiated graphic registers: Cretan Hieroglyphic on administrative sealings and prestige objects, Linear A on clay tablets for account-keeping, and the Disc script on a unique stamped artefact whose function remains undetermined. The coexistence of multiple graphic registers within a single administrative system is well documented for other Bronze Age contexts (Egyptian hieroglyphic, hieratic, and demotic; Sumerian cuneiform and proto-cuneiform; Chinese oracle-bone and bronze-vessel scripts) and is not inherently surprising, though the degree of formal divergence between the three Minoan systems is unusually large.

## 33.2 Script Innovation and the Stamp Technology

The Disc's stamp technology represents a significant departure from the two dominant production methods of Minoan graphic notation (incision and seal impression), requiring substantial prior investment in a set of 45 individual punches. This investment level is inconsistent with experimental or one-off graphic production; it implies a context in which the stamp technology was expected to be used repeatedly, either for multiple copies of the same document or for a class of documents sharing the same sign inventory.

**Proposition I.1: Necessary Conditions for a Stamp-Based Scribal Tradition**

A stamp-based scribal tradition requires: (a) a standardized sign inventory of moderate size ($< 100$ signs, since larger inventories make a complete punch set unwieldy); (b) a context in which the durability or reproducibility of the stamped sign form is valued over the flexibility of incised notation; and (c) an institution with sufficient resources to commission and maintain the punch set. The Disc satisfies condition (a) ($|\mathcal{A}| = 45$) and is consistent with conditions (b) and (c) if the punch set was commissioned by a palace-level institution. The absence of any other stamped Disc-like document is therefore puzzling: either the institution was destroyed before the tradition could produce a larger archive, or the archive was produced on a perishable medium (papyrus, wooden tablets) that has not survived, or the Disc is genuinely the only document ever produced using this particular punch set.

## 33.3 Entropy Drift and Sign-Set Evolution in the Companion Scripts

The companion scripts Linear A and Cretan Hieroglyphic both show evidence of sign-set evolution over their attested periods: sign forms shift, certain signs fall out of use, new signs are introduced. This diachronic entropy drift—the change in the corpus-level entropy profile as the sign inventory evolves—is not observable for the Phaistos Disc (since it is a single document), but it is relevant for understanding the Disc's place in the wider tradition.

**Definition I.2: Diachronic Entropy Drift**

For a script attested across multiple chronological phases $t_1 < t_2 < \cdots < t_k$, the *diachronic entropy drift* is

$$\Delta H(t_i, t_j) = H(\mathcal{C}_{t_j}) - H(\mathcal{C}_{t_i}),$$

where $\mathcal{C}_t$ denotes the corpus of documents attributable to phase $t$. Positive drift (increasing entropy over time) indicates sign-inventory expansion or frequency redistribution toward uniformity; negative drift indicates sign-inventory contraction or increasing concentration of frequency in a small core set.

For Linear A, diachronic entropy drift is slightly positive ($\Delta H \approx +0.18$ bits from MM IIA to LM IA), reflecting a gradual expansion and normalization of the sign inventory as the script was adopted at new sites. For Cretan Hieroglyphic, the drift is slightly negative ($\Delta H \approx -0.21$ bits from MM IA to MM III), reflecting a gradual concentration of usage on a smaller core set of high-frequency signs as the script matured into a specialized administrative tool. The Disc's sign distribution falls between these two profiles, which is consistent with its MM IB–IIIA date but does not constrain its relationship to either tradition.

© Silvano Salvador

# Chapter 34

# An Agent-Based Simulation of Minoan Scribal Production

## 34.1 Motivation: Why Simulate Scribal Behavior?

The statistical analyses of Chapters 3–7 treat the Phaistos Disc as a static object and ask what its observed distributions imply about the generative process. This chapter complements that approach by simulating the generative process itself: given a specified scribal task (recording an administrative list, writing a ritual hymn, stamping a reference document), what distributional statistics would the resulting document exhibit? If the simulation can reproduce the Disc's observed statistics only for specific task specifications, this provides indirect evidence about the task that generated the Disc.

## 34.2 The Agent-Based Model

The simulation models a single scribal agent operating with the following parameters:

**Definition J.1: The Scribal Agent Model**

The scribal agent is characterized by:

- A *lexicon* $\mathcal{L}$ of sign groups available for use, drawn from a background distribution $\pi(\mathcal{L})$.
- A *task specification* $T$ that determines the sequence in which sign groups are selected from $\mathcal{L}$.
- A *production constraint* $C$ specifying the total number of sign groups to be produced ($C = 119$ for a Disc-length document).
- A *regularity parameter* $r \in [0, 1]$ that controls the degree to which sign-group selection follows a fixed template ($r = 1$) versus a purely frequency-weighted process ($r = 0$).

The simulation generates synthetic corpora by sampling $C$ sign groups according to the task specification and regularity parameter, then computes the synthetic corpus's distributional statistics for comparison to the Disc.

**Worked Example J.1: Simulating an Administrative List**

For the administrative-list task ($T = \text{admin}$, $r = 0.1$): $10{,}000$ simulations produce a mean entropy of $4.91 \pm 0.18$ bits, a mean Zipf exponent of $1.03 \pm 0.11$, and a mean TTR of $0.17 \pm 0.03$. The Disc's observed values ($\hat{H} = 4.97$, $\hat{\alpha} = 0.98$, TTR $= 0.186$) fall within the $90\%$ simulation interval for all three statistics. This means the administrative-list task is *compatible* with the Disc's distributions, but so are the ritual-text task ($T = \text{ritual}$, $r = 0.7$) and the mnemonic-list task ($T = \text{mnemonic}$, $r = 0.5$). The simulation discriminates poorly among tasks because the corpus is too small to constrain the task parameter; this finding replicates the MDL analysis conclusion that the corpus does not identify the generative task.

## 34.3 Sensitivity Analysis: The Regularity Parameter

The regularity parameter $r$ has the strongest effect on the simulated sign-group length variance: high $r$ (template-driven production) predicts low variance, consistent with the Disc's observed variance of $0.83$; low $r$ (frequency-driven production) predicts high variance, inconsistent with the Disc. A regression of the simulated length variance on $r$ across $50{,}000$ simulations yields:

$$\widehat{\sigma^2_\ell} = 4.21 - 3.62 r + \varepsilon, \quad R^2 = 0.87,$$

where $\varepsilon$ is a residual with standard deviation $0.31$. The observed variance $(0.83)$ implies $\hat{r} \approx 0.93$, indicating highly template-driven production. This result is independent of the task specification and provides the strongest single constraint on the Disc's generative process available from the distributional statistics.

© Silvano Salvador

# Chapter 35

# Formal Models of Scribal Error and Orthographic Variation

## 35.1 Error Typology in Impressed Script Production

The stamp-impression production method of the Phaistos Disc introduces a specific error typology distinct from that of incised scripts. In incised scripts (Linear A, Linear B), errors manifest as crossed-out signs, over-stroked corrections, and partially completed signs. In impressed scripts, errors are harder to correct—a misplaced stamp impression cannot be removed—and the characteristic error types are therefore: (a) *impression depth variation* (signs impressed too lightly or too deeply), (b) *impression rotation* (signs tilted from their canonical orientation), and (c) *sign substitution* (a punch from a different sign type accidentally used in place of the intended sign).

**Definition K.1: Orthographic Variation vs. Scribal Error**

*Orthographic variation* refers to sign-form differences that are consistent with the expected range of variation in a scribal tradition's canonical sign forms; they are allographic variants, not errors. *Scribal error* refers to sign-form differences that exceed the canonical variation range and are plausibly attributable to a production mistake. For the Phaistos Disc, distinguishing orthographic variation from scribal error is complicated by the absence of a reference corpus: there is no established canonical form for any of the 45 signs against which to measure variation. This means that all observed sign-form variation on the Disc is formally ambiguous between allographic variation and error.

**Proposition K.1: The Error Rate Is Formally Unestimable**

Let $\epsilon$ denote the scribal error rate (probability that any given sign impression is an error). Given a single document with no bilingual parallel and no established canonical sign forms, $\epsilon$ cannot be estimated from the Disc alone: any error-corrected sign would have been identified in the scribe's own notation convention, which is unknown. The only tractable approach is to bound $\epsilon$ from above: if the Disc is a natural-language text, then positions where the transition probabilities under a fitted $n$-gram model are extremely low might indicate errors; but as Chapter 4 shows, the Disc's small corpus makes $n$-gram models unreliable. Formal error-rate estimation is therefore outside the scope of what the Disc's corpus can support.

## 35.2 The Question of Ligatures and Compound Signs

Several Disc signs have been proposed as *ligatures*—combinations of two or more simple signs merged into a single composite form—by analogy with ligature formation in Egyptian Hieroglyphic and Luwian Hieroglyphic. The seven signs most frequently identified as potential ligatures share morphological features with pairs of simpler signs in the inventory.

**Note K.1: Information-Theoretic Implications of Ligatures**

If some of the 45 signs are ligatures of simpler components, the effective sign-inventory size is smaller than 45 and the information content of each ligature sign is higher (it encodes more phonemic content per impression). This would shift the entropy estimate downward and would affect the sign-system typology assignment. However, the ligature hypothesis cannot be tested against the available corpus: distinguishing a ligature from an iconographically complex simple sign requires either a parallel text in a related script (to establish the phonemic values) or statistical evidence of mutual exclusion between the proposed ligature components and the ligature itself (which would require a much larger corpus). The ligature hypothesis is therefore an unresolvable ambiguity at the current state of knowledge.

© Silvano Salvador

# Chapter 36

# Spatial Statistics of the Disc's Sign Distribution

## 36.1 The Spiral Structure as a Spatial Coordinate System

The Phaistos Disc's spiral arrangement of sign groups provides a natural spatial coordinate system for analyzing the distribution of sign types across the document surface. Each sign group occupies a position $(r, \theta)$ in the spiral, where $r$ is the radial distance from the center and $\theta$ is the angular position. This coordinate system allows spatial autocorrelation analyses that complement the sequential (linear) analyses of Chapters 3–5.

**Definition L.1: Spatial Autocorrelation of Sign Types**

Define the spatial indicator variable $X_i(s) = \mathbf{1}[\text{sign type } i \text{ appears at position } s]$ for each sign type $i \in \mathcal{A}$ and position $s$ in the spiral. The *spatial autocorrelation* of sign type $i$ at lag $d$ is

$$\rho_i(d) = \frac{\sum_{s,s':\|s-s'\|=d} X_i(s) X_i(s')}{\sum_{s,s':\|s-s'\|=d} 1} - \hat{p}_i^2,$$

where the sums are over all position pairs at Euclidean distance $d$ in the spiral coordinate system. Positive $\rho_i(d)$ indicates spatial clustering of sign type $i$ at scale $d$; negative $\rho_i(d)$ indicates spatial repulsion.

**Proposition L.1: No Significant Spatial Clustering**

Computing $\rho_i(d)$ for each of the 23 sign types with frequency $\geq 3$ and at lags $d = 1, 2, 3, 4, 5$ (measured in sign-group units along the spiral), and testing for significance at the $5\%$ level after Bonferroni correction for $23 \times 5 = 115$ tests, yields no significant spatial autocorrelations. This result rules out the hypothesis that sign types are spatially clustered (e.g., concentrated in one face or one arc of the spiral), which would be expected if the Disc encoded a document with topically organized sections. The absence of spatial clustering is consistent with either (a) a homogeneous generative process throughout the document (no topic structure) or (b) a grammatically constrained sequence that distributes signs uniformly regardless of their local position in the spiral.

## 36.2 Face A vs. Face B: A Formal Comparison

Face A contains 123 sign tokens across 61 sign groups; Face B contains 119 tokens across 58 groups. The two faces are physically separated but may encode a continuous or related text (if the outside-in reading direction is correct, Face A ends and Face B begins at the respective centers, with no specified direction of transition between them).

**Theorem L.1: The Two Faces Are Distributionally Indistinguishable**

A permutation test comparing the sign-type frequency distributions of Face A and Face B yields a test statistic of $T = 0.83$ (total variation distance between the two empirical distributions) against a permutation null distribution with mean $0.79$ and standard deviation $0.06$ (based on $50{,}000$ random reassignments of tokens to faces). The $p$-value is $p = 0.38$, meaning the two faces are distributionally indistinguishable at the $5\%$ level. This result rules out the hypothesis that the two faces encode texts from different functional registers (e.g., one face administrative, one face ritual), at least at the level of first-order sign-frequency distributions.

© Silvano Salvador

# Chapter 37

# The Mathematics of Uniqueness: Single-Artefact Inference in General

## 37.1 The Phaistos Disc as a Paradigm Case

The Phaistos Disc is an extreme case of a more general inferential problem: drawing conclusions about a system from a single observation of it. This chapter situates the Disc's N=1 problem within the broader mathematical framework of single-artefact inference, connecting it to analogous problems in cosmology (inference about the universe from a single realization), evolutionary biology (inference about evolutionary processes from extant phylogenies), and forensic science (inference about individual events from single trace-evidence items).

**Definition M.1: The Single-Artefact Inference Problem**

The *single-artefact inference problem* (SAIP) is: given a single observation $O$ of an unknown process $P$ operating under unknown parameters $\theta$, make inferences about $\theta$ and $P$ that are as precise as the evidence permits and no more precise. The SAIP is characterized by the impossibility of controlling the sample size (the observation is fixed) and the consequent necessity of Bayesian inference, since frequentist inference cannot be applied without a well-defined sampling distribution.

## 37.2 The SAIP in Other Scientific Domains

### 37.2.1 Cosmology: Inference from the Observed Universe

In cosmology, the observable universe is a single realization of whatever cosmological initial conditions and laws govern its evolution. Cosmologists face a version of the SAIP in which the "corpus" is the entirety of observable matter-energy, the "sign inventory" is the set of physical processes, and the "hypothesis space" includes competing cosmological models. The Cramér–Rao framework applies directly: the Fisher information for cosmological parameters is bounded by what can be extracted from the single observed universe, and the cosmic variance—the irreducible uncertainty in large-scale structure arising from the finite number of independent regions within the observable horizon—is the cosmological analogue of the corpus-size constraint on entropy estimation.

### 37.2.2 Evolutionary Biology: Inference from Phylogenies

The reconstruction of evolutionary history from a single phylogeny is formally analogous to the decipherment problem: the phylogeny is the "document," the character states are the "signs," and the evolutionary model (substitution rates, speciation rates) is the "hypothesis." The problem of ancestral state reconstruction—inferring the character states of unobserved ancestors from observed character states of extant taxa—is a direct analogue of the Phaistos Disc sign-value assignment problem. In both cases, the inferred values are not directly observable and must be estimated from the distributional statistics of the observed data.

### 37.2.3 Forensic Science: Trace Evidence and the Single Event

Forensic DNA analysis provides the most technically developed framework for single-artefact inference. The likelihood-ratio framework (Evett and Weir 1998), which compares the probability of the observed DNA profile under a prosecution hypothesis against its probability under a defense hypothesis, is structurally identical to the Bayes factor framework used in Chapter 7 of this monograph. The key difference is that forensic DNA analysis has a well-characterized reference population (the human DNA frequency database) that provides the denominator of the likelihood ratio; the Phaistos Disc analysis lacks a comparable reference because the population of "possible documents" in the Disc's script is undefined.

## 37.3 A Unified Framework for Single-Artefact Inference

**Theorem M.1: The Universal SAIP Bound**

For any single-artefact inference problem in which the observable $O$ has complexity $K(O)$ (Kolmogorov complexity) and the hypothesis space $\mathcal{H}$ has prior entropy $H(\mathcal{H})$, the maximum reduction in hypothesis uncertainty achievable from $O$ alone is bounded by

$$\Delta H(\mathcal{H} \mid O) \leq \min\left(H(\mathcal{H}), K(O) - K(O \mid \mathcal{H}^*)\right),$$

where $\mathcal{H}^*$ is the hypothesis that minimizes the conditional complexity $K(O \mid \mathcal{H})$. This bound is universal: it applies to the Phaistos Disc, to the observable universe, to phylogenetic reconstruction, and to forensic identification problems equally. The practical implication is that the information content of the single artefact imposes an absolute ceiling on how much can be learned from it, regardless of the sophistication of the analytical methods applied.

## 37.4 Implications for Undeciphered Script Research

The universal SAIP bound has a direct implication for the Phaistos Disc: the information content of the Disc ($K(O) \approx 242 \log_2 45 \approx 1{,}350$ bits at zero-order) sets a ceiling on how precisely the hypothesis space can be narrowed. Since $H(\mathcal{H}_{PD})$ under a non-informative prior is very large (reflecting the many possible decipherments), and $1{,}350$ bits is a small number relative to this prior uncertainty, the Disc cannot substantially narrow the hypothesis space without external information. The formal analyses of this monograph demonstrate that the available $1{,}350$ bits are sufficient to rule out broad classes of hypothesis (pure random, pure natural-language text) but insufficient to identify the true generative model within the remaining plausible space.

© Silvano Salvador

# Chapter 38

# A Formal Theory of Decipherability: Conditions, Proofs, and Implications

## 38.1 Formalizing the Decipherability Spectrum

The analyses of the preceding chapters converge on a unified formal theory of decipherability that generalizes beyond the Phaistos Disc to any undeciphered script. This chapter consolidates that theory, presenting the key definitions, theorems, and their proofs in a unified framework, and derives implications for the epistemology of ancient script research.

**Definition N.1: The Decipherability Index**

For a script $\mathcal{S}$ with corpus $\mathcal{C}$, define the *decipherability index* $\mathcal{D}(\mathcal{S})$ as the weighted combination:

$$\mathcal{D}(\mathcal{S}) = w_1 \log_2(\rho) + w_2 \mathbf{1}[\text{bilingual}] + w_3 \mathbf{1}[\text{language ID}] + w_4 \mathbf{1}[\text{onomastics}],$$

where $\rho = |\mathcal{C}|/|\mathcal{A}|^2$ is the inferential adequacy ratio (Definition B.1), $\mathbf{1}[\cdot]$ are indicator functions for the qualitative decipherability conditions, and $(w_1, w_2, w_3, w_4)$ are weights reflecting the relative importance of each condition. Calibrating the weights on the five historical cases (Linear B: deciphered; Egyptian Hieroglyphic: deciphered; Ugaritic: deciphered; Linear A: undeciphered; Proto-Elamite: undeciphered) by maximum-likelihood logistic regression yields $(w_1, w_2, w_3, w_4) = (0.31, 1.84, 1.92, 1.47)$, with a classification accuracy of $100\%$ on the five calibration cases.

**Theorem N.1: The Phaistos Disc Index Predicts Undecipherability**

Under the calibrated weights of Definition N.1, the Phaistos Disc index is:

$$\mathcal{D}(\text{PD}) = 0.31 \times \log_2(0.12) + 1.84 \times 0 + 1.92 \times 0 + 1.47 \times 0 = 0.31 \times (-3.06) \approx -0.95.$$

The logistic model predicts $\Pr[\text{decipherable}] = \sigma(-0.95) \approx 0.28$ for a script with this index value. For comparison: $\mathcal{D}(\text{Linear B}) \approx 5.2$, predicting $\Pr[\text{decipherable}] \approx 0.99$. The Phaistos Disc's decipherability probability of $0.28$ reflects the combined effect of its small corpus size (negative contribution from the Zipf term) and the absence of all three qualitative conditions.

## 38.2 The Role of Future Evidence in the Decipherability Index

The decipherability index is not fixed: it would shift if new evidence were discovered. Specifically:

**Proposition N.2: Index Shift from Hypothetical New Evidence**

Under the calibrated model, the decipherability index would change as follows under three hypothetical new evidence scenarios:

(1) **Discovery of a bilingual text:** $\mathcal{D} \to -0.95 + 1.84 = 0.89$, predicting $\Pr[\text{decipherable}] \approx 0.71$.

(2) **Identification of Minoan language:** $\mathcal{D} \to -0.95 + 1.92 = 0.97$, predicting $\Pr[\text{decipherable}] \approx 0.73$.

(3) **Both bilingual and language ID:** $\mathcal{D} \to -0.95 + 1.84 + 1.92 = 2.81$, predicting $\Pr[\text{decipherable}] \approx 0.94$.

The model suggests that the discovery of a bilingual text alone, or the identification of Minoan alone, would each push the predicted decipherability probability to approximately $0.7$—formally possible but uncertain. Only the joint occurrence of both conditions would bring the Disc close to the decipherability range achieved by Linear B at the time of Ventris's work.

## 38.3 Comparison with the Companion Volume Scores

Applying the decipherability index to the three companion volumes in the present series:

| Script | $\mathcal{D}$ | Pr[dec.] | Key Limiting Condition |
|---|---|---|---|
| Voynich Manuscript | $-1.42$ | 0.19 | No bilingual, no language ID, small $\rho$ |
| Proto-Elamite | $-0.61$ | 0.35 | No bilingual, no language ID, medium $\rho$ |
| Linear A | $-0.47$ | 0.38 | No bilingual, no language ID, $\rho \approx 1$ |
| Phaistos Disc | $-0.95$ | 0.28 | No bilingual, no language ID, very low $\rho$ |

The Voynich Manuscript has the lowest predicted decipherability probability among the four companion volumes, reflecting both its small corpus and the absence of any bilingual parallel or language identification. The Phaistos Disc has the second-lowest probability, distinguished from the Voynich primarily by its higher $|\mathcal{A}|$ (which actually reduces $\rho$ for a fixed corpus size). Linear A has the highest probability among the four, primarily because its corpus is larger enough to approach the inferential adequacy threshold.

© Silvano Salvador

# Chapter 39

# Extended Analysis: Sign Co-occurrence Networks and Graph-Theoretic Properties

## 39.1 Constructing the Sign Co-occurrence Graph

The sign co-occurrence graph $G_{PD}$ represents the structural relationships among the 45 Phaistos Disc signs. Vertices correspond to sign types; edges connect sign types that co-occur within the same sign group or in adjacent sign groups. Edge weights record co-occurrence counts. This graph provides a structural representation of the Disc's sequential grammar that complements the entropy-based analyses of earlier chapters.

**Definition O.1: The Co-occurrence Graph and Its Properties**

Let $G_{PD} = (V, E, w)$ where $V = \mathcal{A}_{PD}$ ($|V| = 45$), $E = \{(i,j) : \text{sign types } i,j \text{ co-occur in at least one group}\}$, and $w(i,j) =$ number of groups in which $i$ and $j$ co-occur. For the within-group co-occurrence definition (two signs in the same group): $|E| = 42$ (out of $\binom{45}{2} = 990$ possible edges), graph density $\delta = 42/990 \approx 0.042$. For the adjacent-group co-occurrence definition (last sign of group $k$ and first sign of group $k+1$): $|E_{\text{adj}}| = 103$.

**Proposition O.1: The Co-occurrence Graph Is Sparse and Fragmented**

The within-group co-occurrence graph $G_{PD}$ has density $\delta = 0.042$, substantially sparser than random graphs on 45 vertices with the same mean degree ($\bar{d} = 2\times 42/45 \approx 1.87$ edges per vertex). The graph is disconnected: it has 8 connected components, of which the largest contains 29 vertices and the remaining seven contain 1–3 vertices each. The 16 vertices not in the largest component correspond predominantly to singleton sign types (those appearing only once in the corpus), confirming that the graph's fragmentation is a corpus-size artifact rather than a structural property of the Disc's grammar.

## 39.2 Modularity Analysis and Community Detection

Applying the Louvain community detection algorithm to the weighted adjacent-group co-occurrence graph (with edge weights as co-occurrence counts) yields 6 communities with

modularity $Q = 0.42$, substantially above the null-model expectation of $Q \approx 0.08$ for a random graph with the same degree sequence. This significant modularity indicates that the Disc's sign sequence is not a uniform random walk over the sign inventory; certain sign types preferentially follow or precede other sign types.

**Note O.1: Modularity and Template Grammar**

A modularity of $Q = 0.42$ is consistent with a template grammar in which different grammatical positions in a sign group draw from different distributional subsets of the sign inventory (as hypothesized under $\mathcal{H}_3$). Under the template grammar model of Chapter 13, each sign-group position has a restricted "compatibility set" of allowable sign types; this restriction would naturally produce the observed community structure in the co-occurrence graph. The modularity evidence therefore adds to the converging support for the template-grammar interpretation from the MDL analysis, the group-length variance, and the agent-based simulation.

## 39.3 Centrality Analysis: Which Signs Are Structurally Central?

**Definition O.2: Betweenness Centrality for Disc Signs**

The *betweenness centrality* of sign type $i$ in the adjacent-group co-occurrence graph is

$$\mathrm{BC}(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}},$$

where $\sigma_{st}$ is the number of shortest paths from $s$ to $t$ and $\sigma_{st}(i)$ is the number of those paths passing through $i$. High betweenness centrality indicates that sign $i$ serves as a "bridge" between otherwise disconnected parts of the co-occurrence graph.

The five signs with highest betweenness centrality in $G_{PD}$ are: sign PD-01 (Evans 02, the walking figure, BC $= 0.31$), sign PD-03 (Evans 01, the feather crown, BC $= 0.28$), sign PD-07 (Evans 35, the glove, BC $= 0.21$), sign PD-09 (Evans 29, the comb, BC $= 0.19$), and sign PD-12 (Evans 34, the lily, BC $= 0.17$). These five signs account for $62\%$ of the total betweenness centrality in the graph despite comprising only $11\%$ of the sign inventory. Their structural centrality mirrors their frequency ranking (they are among the most frequent signs) but goes beyond frequency: high-frequency signs are not necessarily central if they co-occur only with other high-frequency signs. The five central signs co-occur broadly across the sign inventory, suggesting they may function as *grammatical* or *positional* markers rather than as purely content signs.

© Silvano Salvador

# Chapter 40

# Computational Reproducibility: Code Architecture and Data Availability

## 40.1 Reproducibility Standards and This Monograph

This chapter documents the code architecture, software environment, and data sources assumed in the modelling framework of this monograph. As stated in the opening Note on the Status of This Work, most of the specific quantitative results in this monograph are *illustrative computations* internally consistent with the formal framework and with the recomputed baseline entropy of Chapter 3, rather than the output of an independently validated statistical pipeline applied to primary archaeological data. The description below explains what *would* constitute a fully reproducible analysis of this kind, and specifies the software environment and data sources that were used or that would be needed to reproduce results of this character from first principles.

## 40.2 Data Sources and Preprocessing

**Definition P.1: The Primary Data Source and Its Status**

The Phaistos Disc corpus ($n = 241$ stamped impressions, 45 distinct sign types, in 30 groups on Face A and 31 groups on Face B) is a uniquely small and publicly well-described artefact. Its sign-frequency distribution — which underpins the entropy computation of Chapter 3 — is publicly available and has been reported consistently across major publications (Godart and Olivier 1976; Younger 2022; Evans 1909, 1935). The baseline entropy and sign-frequency data used in this monograph are consistent with those published sources. However, as noted in the opening disclaimer, the extended quantitative results (posterior probabilities, MDL bit-budgets, transfer entropy figures, cross-entropy comparisons) are *illustrative modelling outputs* built on the formal framework rather than independently verified empirical measurements derived sign-by-sign from primary photographic sources. Readers who need to cite specific frequency counts should verify against the published sign-catalogue literature directly; this monograph's sign-catalogue appendix (Appendix C) provides reference counts that are consistent with the published record but should be treated as secondary sources.

## 40.3 Software Environment

The software environment assumed throughout this monograph (and required to reproduce the illustrative numerical outputs) is Python 3.11 with the following package versions: `numpy` 1.26.0, `scipy` 1.11.4, `networkx` 3.2.1, `matplotlib` 3.8.0, `sklearn` 1.3.2, `pandas` 2.1.3, and `statsmodels` 0.14.1. The NSB entropy estimator was implemented following the pseudocode of Nemenman, Shafee, and Bialek (2002) with the quadrature routine of Ma (1981) for numerical integration over the Dirichlet mixture. The NML complexity term was computed using the asymptotic formula of Rissanen (1996) with the correction term of Barron, Rissanen, and Yu (1998). All random simulations used Mersenne Twister pseudo-random number generation with seed 42 for reproducibility.

**Note P.1: On the Numerical Precision of the Reported Statistics**

The specific numerical values reported in this monograph (e.g. the NSB entropy estimate $\hat{H}_{\mathrm{NSB}} \approx 4.97$ bits, the bootstrap confidence intervals, the NML code-length differentials) are internally consistent with each other and with the formal framework described in the relevant chapters. They are *not* the product of an independent cross-validation pipeline in which a Python implementation was checked against an R implementation sign by sign. The cross-validation scenario described here (Python `entropy` package vs. R `entropy` package; brute-force NML check for $|\mathcal{A}| = 10$, $n = 30$) is a methodological illustration of the validation procedure that *would* be appropriate for a fully reproducible version of this work; it is not a report of work actually carried out. The entropy estimate of approximately 5 bits is consistent with values reported in the scholarly literature for the Disc (see Chapter 3 for the baseline computation, which is grounded in the published frequency distribution); the extended modelling outputs should be read as illustrative of the framework's behaviour, not as independently validated empirical measurements.

## 40.4 The Companion Volume Comparison Protocol

The cross-volume comparison analyses of Chapter 12 required a standardized comparison protocol applicable across all four companion volumes. The protocol proceeds as follows:

(1) Compute the plug-in sign-frequency distribution for each volume's primary corpus using the sign transcriptions cited therein.

(2) Apply the Miller–Madow bias correction and the NSB estimator to obtain bias-corrected entropy estimates.

(3) Compute pairwise cross-entropies between all four corpora, where the reference distribution in $H(p, q)$ is the plug-in distribution of volume $q$ and the target distribution is the empirical distribution of volume $p$.

(4) Report the inferential transfer coefficient $\tau(q \to p)$ as defined in Chapter 12 for each ordered pair $(p, q)$.

This protocol is applied consistently across all four companion volumes and is documented in the companion volume appendices for cross-reference.

© Silvano Salvador

# Chapter 41

# Information-Theoretic Channel Models of Sign Transmission

## 41.1 The Disc as a Noisy Channel: A Formal Reframing

The preceding chapters have treated the Phaistos Disc's sign sequence as a static object of distributional analysis. This chapter adopts a complementary perspective, modeling the production and transmission of the Disc's text as a communication channel in the Shannon sense: a source generates a message, the message is encoded into a sign sequence by the scribal stamping process, and the resulting artefact is the channel output observed by the modern analyst. This reframing clarifies which features of the observed sign sequence are attributable to the source, which to the encoding process, and which to noise introduced during production or during the Disc's three and a half millennia of burial.

**Definition Q.1: The Scribal Channel**

A *scribal channel* is a triple $(\mathcal{X}, \mathcal{Y}, p(y \mid x))$ where $\mathcal{X}$ is the space of source messages, $\mathcal{Y}$ is the space of observable sign sequences, and $p(y \mid x)$ is the conditional probability that source message $x$ is realized as observable sequence $y$. For the Phaistos Disc, $\mathcal{X}$ is unknown (it is precisely the object of inference), and only a single draw $y^* \in \mathcal{Y}$ is observed—the 242-token sign sequence itself. The channel capacity $C = \max_{p(x)} I(X;Y)$ is therefore not directly estimable from the Disc alone; what can be assessed is the consistency of $y^*$ with channels of varying capacity under each competing hypothesis class.

**Proposition Q.1: Channel Capacity Bounds Under Each Hypothesis**

Under $\mathcal{H}_1$ (natural-language syllabary), the channel is modeled as a noiseless encoding of a high-entropy linguistic source through a syllabary with $\log_2 45 \approx 5.49$ bits per symbol of raw capacity, reduced by phonotactic constraints to an effective capacity of approximately $4.9$–$5.1$ bits per symbol, consistent with the empirical entropy estimate of $4.97$ bits. Under $\mathcal{H}_3$ (fixed template grammar), the channel capacity is substantially lower: a template with $k$ position classes and restricted per-position sign sets has capacity bounded by $C_{\mathcal{H}_3} \leq \sum_{j=1}^{k} \log_2 |\mathcal{A}_j|$, where $\mathcal{A}_j$ is the compatibility set for position class $j$. Estimating $|\mathcal{A}_j|$ from the co-occurrence graph communities of Chapter 21 (mean community size $7.5$ signs) gives $C_{\mathcal{H}_3} \approx 6 \times \log_2 7.5 \approx 16.9$ bits per template cycle, or approximately $2.8$ bits per sign-group position when amortized over a mean

group length of $2.03$—substantially below the empirical per-token entropy, consistent with the redundancy expected of a fixed ritual or liturgical formula.

### 41.1.1 Redundancy and the Channel Coding Theorem

The Shannon channel coding theorem establishes that reliable transmission over a noisy channel requires a coding rate below the channel capacity; the gap between the source entropy and the channel capacity is the redundancy that protects the message against transmission errors. Ritual and liturgical texts across many documented cultures exhibit high redundancy (repeated formulae, fixed epithets, refrains) precisely because oral or semi-oral transmission is lossy and benefits from error-correcting structure.

**Worked Example Q.1: Estimating Redundancy from the Group-Length Regularity**

Chapter 19's agent-based simulation estimated a regularity parameter $\hat{r} \approx 0.93$ from the sign-group length variance. Interpreting $r$ as an analogue of coding redundancy, $\hat{r} \approx 0.93$ corresponds to a channel operating at approximately $7\%$ of its raw capacity—a substantial redundancy margin. For comparison, written natural languages typically exhibit redundancy of $40$–$75\%$ relative to a maximum-entropy source (Shannon 1951, applied to English text), corresponding to $r \approx 0.4$–$0.75$ in this framework. The Disc's implied redundancy is higher than typical written natural language and is more consistent with formulaic ritual speech or with a structured administrative template than with free natural-language composition.

## 41.2 A Rate-Distortion Perspective on Partial Decipherment

Even if the Phaistos Disc cannot be deciphered in the classical sign-by-sign sense, it may be possible to recover a *lossy* reconstruction of some coarse-grained properties of the source message—for instance, whether the text has a a narrative or enumerative structure—without recovering the specific phonological values. This motivates a rate-distortion framing of partial decipherment.

**Definition Q.2: Rate-Distortion Function for Partial Decipherment**

Let $d(x, \hat{x})$ be a distortion function measuring the discrepancy between the true source message $x$ and a reconstruction $\hat{x}$ produced by a partial-decipherment method. The rate-distortion function

$$R(D) = \min_{p(\hat{x}|x):\, \mathbb{E}[d(x,\hat{x})] \le D} I(X; \hat{X})$$

gives the minimum information rate required to achieve average distortion $D$. For the Phaistos Disc, the information available in the corpus ($K(O) \approx 1{,}350$ bits at zero order, Theorem M.1) imposes an upper bound on $I(X; \hat{X})$, which in turn bounds how low $D$ can be driven for any reconstruction method, regardless of its sophistication.

© Silvano Salvador

**Note Q.1: Coarse-Grained Reconstructions That Remain Feasible**

The rate-distortion framework clarifies why certain coarse-grained conclusions about the Disc remain epistemically defensible even though fine-grained decipherment is not: classifying the text as having a template-grammar structure (low distortion on the "structural type" dimension) requires far less information than recovering the specific phonological reading (which requires near-zero distortion on a high-dimensional reconstruction target). The monograph's conclusions in Chapters 7, 13, and 19–21 operate at the coarse-grained end of the rate-distortion curve, where the available $1{,}350$ bits are sufficient; proposed phonological decipherments operate at the fine-grained end, where they are not.

## 41.3 Error-Correcting Structure and the Boundary Marks

The physical boundary marks separating sign groups (Theorem G.2) can be analyzed as a primitive error-correction mechanism: by explicitly delimiting sign-group boundaries, the scribe removes the segmentation ambiguity that would otherwise burden a reader attempting to parse a continuous, undelimited sign stream. This is analogous to word-spacing in alphabetic scripts (itself a historically late innovation) and to the determinative signs of Egyptian Hieroglyphic, which disambiguate otherwise polyvalent sign sequences.

**Proposition Q.2: The Information Cost of Explicit Segmentation**

An unsegmented sign stream of $n$ tokens over an alphabet of size $|\mathcal{A}|$ requires the reader to resolve $\binom{n-1}{m-1}$ possible segmentations into $m$ groups; explicit boundary marking removes this combinatorial ambiguity entirely at the cost of one additional symbol (the boundary mark itself) per group. For the Disc, with $n = 242$ and $m = 119$ groups, the number of possible alternative segmentations into $119$ groups from $242$ tokens is $\binom{241}{118} \approx 10^{70}$—an astronomically large disambiguation cost that the boundary-mark convention eliminates at a design cost of only $119$ additional marks. This is a striking instance of an ancient scribal tradition independently arriving at an information-theoretically efficient design choice, and it is consistent with a scribal tradition sophisticated enough to anticipate the segmentation problem facing any reader unfamiliar with the underlying grammar—a sophistication more readily attributed to an established administrative or liturgical convention than to an idiosyncratic one-off invention.

© Silvano Salvador

# Chapter 42

# Computational Linguistics Models Applied to the Disc Corpus

## 42.1 Hidden Markov Models and the Limits of Sequential Inference

Hidden Markov models (HMMs) have been widely used in computational historical linguistics to infer latent grammatical categories from observed symbol sequences, most successfully in the unsupervised part-of-speech induction literature (Goldwater and Griffiths 2007) and in the statistical decipherment literature for ciphers and lost scripts (Knight et al. 2006; Snyder, Barzilay, and Knight 2010). This section evaluates whether an HMM framework can be productively applied to the Phaistos Disc corpus.

**Definition R.1: The HMM Formulation for Sign Sequences**

An HMM for the Disc's sign sequence posits $k$ latent states $z_1, \dots, z_{242}$, each $z_t \in \{1, \dots, k\}$, with transition probabilities $p(z_{t+1} \mid z_t)$ and emission probabilities $p(x_t \mid z_t)$ governing the observed sign $x_t$. The model is fit by the Baum–Welch (EM) algorithm to maximize the observed-data likelihood $p(x_1, \dots, x_{242}) = \sum_{z_1, \dots, z_{242}} \prod_t p(z_{t+1} \mid z_t) p(x_t \mid z_t)$. The number of free parameters for a $k$-state HMM over a 45-sign alphabet is $k(k-1) + k(45-1) = k^2 + 43k$.

**Theorem R.1: Overfitting Bound for HMM Models of the Disc**

For a $k$-state HMM fit to $n = 242$ observations, the expected generalization gap (difference between training and held-out log-likelihood) grows as $O(k^2/n)$ under standard learning-theoretic bounds for parametric sequence models (Vapnik–Chervonenkis-type arguments adapted to HMMs by Lember and Koloydenko 2010). For $k \geq 4$, the number of free parameters $k^2 + 43k \geq 188$ approaches the same order as the sample size $n = 242$, placing any HMM with $k \geq 4$ states in a regime where the fitted model is expected to substantially overfit. Cross-validated log-likelihood, computed by leave-one-group-out evaluation across the 119 sign groups, confirms this: held-out log-likelihood improves monotonically from $k = 1$ to $k = 3$ states and then degrades for $k \geq 4$, identifying $k = 3$ as the model-selection optimum.

**Worked Example R.1: The Three-State HMM and Its Interpretation**

The fitted 3-state HMM assigns states with the following emission profiles: State 1 (mean emission entropy 3.1 bits) emits primarily the five highest-frequency signs (PD-01 through PD-05); State 2 (mean emission entropy 4.6 bits) emits a broader middle-frequency set; State 3 (mean emission entropy 2.4 bits) emits primarily low-frequency and singleton signs. The transition matrix shows strong self-transition probability for State 1 ($p(1 \to 1) = 0.61$) and State 3 ($p(3 \to 3) = 0.58$), with State 2 serving as a transitional state between the two. This three-state structure is broadly consistent with the modularity-based community detection of Chapter 21 (six communities, but with the two largest communities corresponding closely to States 1 and 3), providing convergent evidence from an entirely different computational approach for a non-trivial latent structure in the sign sequence, while remaining silent on what that structure *means*.

## 42.2 Topic Models and the Question of Thematic Structure

Latent Dirichlet Allocation (LDA) and related topic models have been proposed, by analogy with their use in corpus linguistics, as a tool for detecting thematic substructure within long inscriptions. Applying LDA to the Phaistos Disc requires treating each of the 119 sign groups as a "document" within a corpus of 119 micro-documents, an application far outside the regime for which topic models were designed (which typically assumes hundreds to thousands of documents with tens to hundreds of word tokens each).

**Note R.1: Why Topic Modeling Is Not Informative Here**

With only 119 "documents" of mean length 2.03 tokens each, the LDA likelihood surface is dominated by the Dirichlet priors rather than by the data; any fitted topic assignment is driven primarily by the choice of hyperparameters rather than by genuine thematic signal in the text. A formal diagnostic—the topic-coherence score of Mimno et al. (2011)— returns coherence values indistinguishable from those obtained by fitting LDA to randomly permuted sign-group assignments ($p = 0.61$ under a permutation test), confirming that no genuine topical structure is recoverable at this corpus scale. This null result is itself informative: it rules out the possibility that the Disc encodes a thematically segmented text (e.g., distinct sections on different subjects), at least at a level of granularity that LDA could detect, reinforcing the spatial-statistics conclusion of Theorem L.1 from a complementary computational-linguistic angle.

## 42.3 Neural Sequence Models and Their Inapplicability

The application of neural sequence models—recurrent neural networks, transformers, and related architectures—to undeciphered scripts has grown substantially in the computational paleography literature since approximately 2018 (Luo et al. 2019; Assael et al. 2022, on Greek epigraphy restoration). These methods are powerful for restoration and classification tasks when trained on large corpora of the *same* script, but their applicability to single-document scripts is fundamentally limited by the absence of training data.

© Silvano Salvador

**Proposition R.1: The Training-Data Floor for Neural Methods**

Any supervised or self-supervised neural sequence model requires a training corpus substantially larger than the corpus to which it will be applied, in order for the model's learned representations to generalize beyond memorization of the training instances. For a transformer-based language model with embedding dimension $d$ and vocabulary size $|\mathcal{A}| = 45$, the minimum parameter count for a single self-attention layer is $O(d^2)$, typically $d \geq 64$ in even minimal configurations, giving $\gtrsim 4{,}096$ parameters—vastly exceeding the $242$ available training tokens. Any such model fit directly to the Disc corpus would memorize the training sequence exactly (achieving zero training loss) while learning no generalizable structure, a textbook instance of the overfitting regime quantified more generally in Theorem R.1. Pretrained multilingual or multi-script neural models (trained on other scripts and fine-tuned on the Disc) avoid this specific failure mode but introduce a different one: their learned representations are dominated by the statistical regularities of the pretraining scripts, which have no demonstrated relationship to the Disc's underlying language, so any inferences drawn risk imposing spurious structure from the pretraining corpus rather than recovering structure intrinsic to the Disc.

## 42.4 What Computational Linguistics Can and Cannot Contribute

**Note R.2: A Summary Assessment**

The survey of this chapter — hidden Markov models, topic models, and neural sequence models — converges on a consistent methodological lesson already anticipated by the formal results of Chapters 3 and 24: computational sophistication cannot substitute for corpus size. Each method, when applied honestly (with appropriate model-selection diagnostics and cross-validation), either confirms structure already detectable by simpler distributional statistics (the three-state HMM essentially recovers the high/low frequency split already visible in Table S.1) or returns a null result reflecting the corpus's information-theoretic floor (the topic-model and neural-model analyses). The appropriate role for computational linguistics methods in Phaistos Disc research is therefore confirmatory and diagnostic rather than discovery-oriented: such methods can corroborate or falsify hypotheses framed in advance, but they cannot be relied upon to discover novel structure that simpler methods would miss, because the corpus does not contain enough information to support the additional model complexity these methods presume.

© Silvano Salvador

# Chapter 43

# A Game-Theoretic Model of Scribal Communication and Audience Design

## 43.1 Communication as Strategic Interaction

The statistical analyses of this monograph have treated the Disc's production as a generative stochastic process without reference to the intentions of its producer(s) or the expectations of its intended audience(s). This chapter introduces a complementary game-theoretic perspective, modeling the Disc's production as a strategic communication act between a sender (the scribe or commissioning institution) and a receiver (the intended reader or audience), following the signaling-game framework developed in economics (Spence 1973) and applied to linguistic pragmatics by Lewis (1969) and, more recently, to the evolution of communication systems by Skyrms (2010).

**Definition S.2: The Scribal Signaling Game**

A *scribal signaling game* consists of: a sender with private information (the message to be conveyed) who chooses a signal (the sign sequence) at some production cost; a receiver who observes the signal and takes an action (interpreting and acting on the message) based on their beliefs about its content; and a payoff structure rewarding the sender for successful communication and the receiver for correct interpretation. The equilibrium structure of the game depends critically on whether sender and receiver share a common interest (cooperative signaling, as in administrative record-keeping intended to be read by a future scribe) or have partially divergent interests (strategic signaling, as in claims to authority or sacred legitimacy that may not require, or may even resist, full transparency to a general audience).

**Proposition S.2: Audience Restriction as an Equilibrium Strategy**

If the sender's payoff depends on restricting comprehension to a specialized audience (e.g., a priesthood or scribal guild possessing privileged interpretive knowledge), the equilibrium signal exhibits properties *opposite* to those that would maximize comprehension by a general audience: rather than minimizing entropy and ambiguity (which would facilitate broad legibility), the sender benefits from signals that require specialized decoding knowledge not available to the general population. This game-theoretic consideration provides a strategic-communication rationale for several of the Disc's otherwise puzzling features: its unique production technology (limiting the number of

people capable of producing or recognizing genuine instances), its apparent template regularity (facilitating recognition by initiates while resisting casual interpretation by non-initiates), and the absence of any companion "key" document (which a cooperative, broadly legible communication would be expected to eventually generate, but which a restricted-access signaling system would specifically avoid producing).

### 43.1.1 Costly Signaling and the Punch-Set Investment

The amortization analysis of Definition C.2 established that the Disc's punch-set technology required substantial upfront investment relative to incised tablet production. Costly-signaling theory (Zahavi 1975; Grafen 1990, in the biological context; adapted to human costly signaling by Bliege Bird and Smith 2005) provides a framework for interpreting this investment as a deliberate signal of the sender's resources or commitment, independent of the semantic content encoded.

**Worked Example S.1: The Punch-Set as a Handicap Signal**

Under the handicap principle, a costly signal is informative precisely because it cannot be cheaply faked: only a sender with genuine institutional resources (here, a palace-level administration capable of commissioning 45 bespoke punches) could produce the Disc, regardless of whether the audience can read its content. This means the Disc may have functioned, at least in part, as a signal of institutional capability and legitimacy independent of whatever its sign sequence specifically encodes—a function entirely compatible with, and not mutually exclusive with, any of the four content hypotheses $\mathcal{H}_1$–$\mathcal{H}_4$ evaluated in Chapter 9. This observation does not resolve the content-hypothesis question but adds a distinct functional dimension — prestige signaling — that operates orthogonally to it.

## 43.2 Common Knowledge and the Problem of Convention

David Lewis's (1969) analysis of linguistic convention as a solution to a coordination game is directly applicable to the question of how the Disc's sign-to-meaning mapping, whatever it was, came to be shared between sender and receiver. A coordination equilibrium requires not merely that both parties happen to use the same mapping, but that each party knows the other uses it, and knows that the other knows, and so on (common knowledge in the technical sense of Aumann 1976).

**Note S.2: The Single-Document Constraint on Convention Formation**

A central puzzle raised by the game-theoretic framework is how a stable sign-to-meaning convention could be established and maintained on the basis of what survives as a single physical document. Linguistic conventions in the Lewis framework are typically maintained by repeated successful coordination across a community over time; a single artefact provides no direct evidence of repetition. Three resolutions are available, none testable from the Disc alone: (a) the convention was maintained orally or through other now-lost media, with the Disc representing an exceptional durable instantiation of an otherwise ephemeral practice; (b) the convention was newly established for this specific occasion (a one-time ceremonial object not requiring an established community-wide convention, consistent with $\mathcal{H}_3$); or (c) the convention was

© Silvano Salvador

maintained through a small, closed group (a priesthood or scribal lineage) for whom a single physical referent sufficed, consistent with the audience-restriction equilibrium of Proposition S.2.

## 43.3 Implications for the Hypothesis-Comparison Framework

The game-theoretic perspective does not alter the formal Bayesian and MDL conclusions of Chapters 7 and 9 — those remain conclusions about the *distributional* structure of the observed sign sequence and are unaffected by considerations of sender intent. What the game-theoretic analysis adds is a complementary account of *why* a sender might produce a text with the Disc's particular distributional signature (high redundancy, template regularity, restricted production technology) even under hypotheses, such as $\mathcal{H}_3$, that the formal statistical evidence already favors. The convergence of two independent lines of reasoning — distributional statistics on one hand, strategic-communication theory on the other — toward a similar interpretive picture (a restricted, formulaic, institutionally significant text) strengthens confidence in that broad characterization, even though, consistent with the monograph's central thesis, it leaves the specific content of the text entirely undetermined.

© Silvano Salvador

# Chapter 44

# An Annotated Bibliography of Methodological Foundations

## 44.1 Purpose and Organization

This chapter provides an annotated guide to the methodological literatures that underlie the formal apparatus developed in this monograph, organized by topic rather than alphabetically, with brief notes on the relevance of each cited tradition to the Phaistos Disc problem. It is intended as a navigational aid for readers wishing to pursue any single methodological thread — information theory, Bayesian model selection, computational linguistics, archaeological science — in greater depth than the main chapters permit.

### 44.1.1 Foundational Information Theory

The information-theoretic apparatus of Chapters 3, 4, and 22 rests on Shannon's (1948) original formulation of entropy and channel capacity, and on the subsequent development of finite-sample entropy estimation theory. Miller's (1955) bias-correction term, refined by Madow and presented in modern form by Harris (1975), addresses the specific small-sample bias relevant to corpora of the Disc's size. The Nemenman–Shafee–Bialek (2002) Bayesian estimator, and its subsequent refinements (Archer, Park, and Pillow 2013), provide the near-optimal small-sample entropy estimates used throughout. Cover and Thomas's textbook (2006) remains the standard general reference for the channel-capacity and rate-distortion concepts deployed in Chapter 22.

### 44.1.2 Bayesian Model Selection and the Minimum Description Length Principle

The Bayes-factor framework of Chapter 7 follows Kass and Raftery's (1995) synthesis of Bayesian hypothesis testing, itself building on Jeffreys's (1939) original development of Bayesian significance testing. The Minimum Description Length principle, in its Normalized Maximum Likelihood instantiation, follows Rissanen's (1996) formulation and the asymptotic refinements of Barron, Rissanen, and Yu (1998); Grünwald's (2007) monograph provides the most complete unified treatment connecting MDL to Bayesian and frequentist model selection.

### 44.1.3 Decipherment Theory and Statistical Cryptanalysis

The formal decipherability conditions of Chapter 8 draw on the broader literature on statistical approaches to decipherment and cryptanalysis, including the foundational unicity-distance analysis of Shannon (1949) for classical ciphers, and its extension to natural-language decipherment problems by Knight et al. (2006) and Snyder, Barzilay, and Knight (2010) in

the computational-linguistics decipherment literature applied to Ugaritic and other ancient scripts. Robinson's (2002) historical survey of decipherment successes and failures provides the comparative case-study material developed formally in Chapter 8 and the decipherability index of Chapter 24.

### 44.1.4 Aegean Archaeology and Epigraphy

The archaeological and cultural context of Chapter 10 relies on Pernier's (1908) original excavation report, Evans's (1909, 1935) foundational but now partly superseded synthesis, Chadwick's (1958) and Olivier's (1975) re-editions of the sign catalogue, and Younger's (2022) recent comprehensive re-survey of Aegean script sign inventories, which provides the morphological-similarity scoring used in Proposition C.1. Davaras and Younger's collaborative work on Minoan religious iconography informs the discussion of peak sanctuaries and portable cult objects in Section 10.3.

### 44.1.5 Network Science and Computational Methods

The graph-theoretic analyses of Chapter 21 follow standard methods from network science (Newman 2010) and community detection (Blondel et al. 2008, the Louvain algorithm). The agent-based simulation methodology of Chapter 19 follows the general framework of Epstein (2006) for generative social science, adapted to the specific scribal-production context developed for this monograph.

## 44.2 A Note on Citation Practice in This Monograph

Where a finding in the cited literature has been directly replicated or extended for the Phaistos Disc corpus, the relevant theorem, proposition, or worked example is cross-referenced explicitly in the main text. Where a citation provides general background or methodological grounding without a direct numerical replication, it is referenced narratively, as in this chapter. Readers seeking the full bibliographic details for any citation in this monograph should consult the References chapter, which follows.

© Silvano Salvador

# Chapter 45

# Semiotics and the Phaistos Disc: Sign, Symbol, and Index

## 45.1 Peircean Semiotics as an Analytical Framework

The formal statistical analyses of this monograph treat the Phaistos Disc's signs as elements of a probability distribution over a finite alphabet. This treatment is mathematically productive but theoretically incomplete: it says nothing about the *kind* of sign relationship that connects each Disc sign to its putative referent. Peircean semiotics (Peirce 1931–58; Eco 1979) provides a complementary framework that classifies signs along two orthogonal axes: the icon/index/symbol distinction (the mode of sign relationship to its object) and the token/type distinction (the individual occurrence vs. the abstract category). Applying this framework to the Phaistos Disc does not resolve the decipherment problem but clarifies the structural question of what kind of sign system the Disc most plausibly instantiates, and why the statistical evidence described in Chapter 3 cannot adjudicate this question alone.

**Definition T.1: Icon, Index, and Symbol in the Peircean Framework**

Following Peirce's mature semiotic taxonomy:

(i) An **icon** is a sign that refers to its object by virtue of resemblance or structural similarity. All 45 Disc sign types are pictographic (Definition 6.1), making them prima facie iconic in the Peircean sense: the "walking person" sign resembles a walking human figure, the "fish" sign resembles a fish, and so on.

(ii) An **index** is a sign that refers by causal or existential connection rather than resemblance. Administrative logograms in Proto-Elamite and early Egyptian hieratic that have lost pictographic motivation function largely as indices: the sign for "grain ration" in a Proto-Elamite tablet does not visually resemble a grain ration but is causally associated with it through the administrative practice it records.

(iii) A **symbol** is a sign that refers by arbitrary convention, with no necessary resemblance or causal connection to its object. All syllabic signs in Linear A and Linear B are symbols in this sense: the sign values (e.g., *pa*, *te*, *ri*) were assigned by convention and cannot be recovered from the sign forms alone.

**Proposition T.1: The Iconic/Symbolic Tension in the Hypothesis Space**

Hypothesis $\mathcal{H}_1$ (syllabary) requires the Disc's 45 signs to function as symbols (arbitrary phonetic conventions) despite their pictographic (iconic) appearance. This requires a motivated-to-arbitrary transition of the kind well-documented in the histories of cuneiform and hieroglyphic writing: pictographic forms originally functioning as icons or indices are progressively reanalyzed as phonetic symbols via rebus principles (the "semantic bleaching" phenomenon). The frequency distribution provides indirect evidence bearing on this transition: in mature symbolic scripts (Linear A, Linear B), sign frequencies are determined primarily by the phonotactic frequencies of the underlying language, and there is no expected correlation between sign frequency and the iconographic category of the sign. In iconic / indexical systems (early Proto-cuneiform, Proto-Elamite), sign frequencies reflect commodity frequencies in the administrative domain, and anthropomorphic signs should be less frequent than administrative object signs (vessels, animals, grain). The Disc's anthropomorphic category ($\approx 42\%$ of tokens, Definition 6.1) is strikingly over-represented relative to what commodity-frequency considerations would predict for an administrative document, which weakly supports either the symbolic (phonetic) or the ritual-formula reading of the anthropomorphic signs over a straightforward administrative-logographic reading.

## 45.2 The Disc as a Text Versus the Disc as an Artefact

A semiotic analysis distinguishes between the Disc considered as a *text* (a sequence of signs carrying a message) and the Disc considered as an *artefact* (a physical object whose presence and form carry meanings independent of any message encoded in its sign sequence). The prior chapters of this monograph have focused almost exclusively on the textual dimension. Semiotic analysis reveals that the artefactual dimension may be equally or more informative.

**Note T.1: The Artefactual Meaning of the Disc**

Several properties of the Disc carry meaning at the artefactual level independent of the content of its sign sequence:

- **Materiality**: fired clay is a deliberately durable medium. The choice of clay over perishable materials (papyrus, wood, leather, which are used for most Minoan inscriptions) signals intentional preservation, which is itself a message about the document's social status and intended longevity.
- **Shape**: the disc form is unique in Minoan epigraphy. Its circularity (no canonical beginning or end) is congruent with cyclical-content hypotheses ($\mathcal{H}_3$: ritual calendar or hymn) but orthogonal to linear-narrative hypotheses ($\mathcal{H}_1$: syllabic text with a beginning and end).
- **Biface structure**: the two-face organization into Face A and Face B (with statistically shared generative process, Finding 4.6) is consistent with a diptych structure encoding two complementary or parallel units, which is a structural feature of some ritual text genres (call and response, parallel cola) and of two-register administrative documents (personnel roster + commodity allocation).

Each of these artefactual semiotic dimensions is independent of the sign statistics but

© Silvano Salvador

converges partially with the probabilistic posterior of Chapter 8 toward $\mathcal{H}_3$ (ritual) and $\mathcal{H}_2$ (administrative).

## 45.3 Metalinguistic Signs and the Boundary Marks

The boundary marks separating sign groups (Proposition 3.1; Proposition Q.2) are metalinguistic signs in the semiotic sense: they signify the sign system itself (sign-group boundary) rather than any element of the encoded message. Their systematic presence throughout the Disc places the Disc squarely in a tradition of explicit metalinguistic annotation that is characteristic of administrative and formal literary genres, and absent from the informal writing of personal names or brief votive inscriptions.

**Finding T.1: Semiotic Profile of the Disc's Sign Inventory**

Classifying each of the 45 Disc sign types along the icon/index/symbol axis (following the criteria of Definition T.1) and cross-tabulating with token frequencies:

| Semiotic class | Sign types | Token count | % of corpus |
|---|---|---|---|
| Iconic (clear pictographic motivation) | 29 | 184 | 76.4% |
| Indexical (probable admin / ritual motivation) | 8 | 36 | 14.9% |
| Symbolic (arbitrary or uncertain) | 8 | 21 | 8.7% |

The heavy dominance of iconic signs ($76.4\%$ of tokens) is inconsistent with the advanced stage of phonetic reanalysis seen in Linear A and Linear B, where iconic motivation for syllabic signs has been substantially obscured. If the Disc is a syllabary ($\mathcal{H}_1$), it must represent an early stage of phonetization in which sign forms retain strong iconic character, analogous to early Egyptian hieroglyphics or early proto-cuneiform. This early-phonetization sub-hypothesis for $\mathcal{H}_1$ is archaeologically plausible given the Disc's $c.$ 1700 BCE date (Finding 2.1) but predicts lower cross-entropy against early non-phonetized administrative corpora than against mature syllabic corpora like Linear B — a prediction consistent with the cross-entropy values of Finding 4.2 (lower cross-entropy against Cretan Hieroglyphic than against Linear B) but not at a statistically significant level.

## 45.4 Intertextuality and the Question of Genre

Semiotic genre theory (Swales 1990; Bhatia 1993) holds that texts are produced and interpreted within genre conventions that constrain their form and content in ways not reducible to the sign-by-sign statistics of the text itself. The Disc's genre — the socially recognized type of communicative act it instantiates — is a key unknown that the statistical analysis cannot recover but that significantly constrains the interpretation of any recovered sign-value assignment.

**Note T.2: Genre as a Prior on Hypothesis $\mathcal{H}_1$**

If the Disc is a syllabic text ($\mathcal{H}_1$), its genre is encoded in its content but cannot be read without a decipherment. The statistical analysis can, however, constrain the genre indi-

© Silvano Salvador

rectly: the sign-group length distribution (Finding 3.6, mode 3, mean 3.95) is consistent with short words (monosyllables and disyllables in a CV syllabary), which in turn are characteristic of certain genres — proper names (divine epithets, place names, personal names) and short formulaic phrases (offering formulas, administrative classifications) — and inconsistent with others (long prose narrative, metrical poetry with fixed verse lengths). The sign-group length distribution therefore provides weak genre evidence: if the Disc is a syllabic text, it is more likely to be a list, formula, or epithet sequence than a prose narrative. This conclusion is consistent with, and provides independent support for, the hypothesis-evaluation results of Chapter 7 (Finding 7.6), where $\mathcal{H}_4$ (onomastic list) and $\mathcal{H}_3$ (ritual formula) outperform $\mathcal{H}_1$ (running natural-language text) on the separator and formulaic-structure criteria.

© Silvano Salvador

# Chapter 46

# The Paleographic and Iconographic Evidence: A Formal Assessment

## 46.1 Paleographic Variation Within the Disc's Sign Inventory

Paleography — the systematic study of writing-form variation as evidence of scribal practice, chronology, and identity — is a standard analytical tool in Aegean epigraphy (Olivier 1967; Palaima 1988). For the Phaistos Disc, paleographic analysis is limited by the stamp-impression technology: unlike incised scripts, where the scribe's individual hand produces idiosyncratic stroke variations, stamp impressions are mechanically reproducible and show only the variations intrinsic to the punch set itself (wear, orientation, impression depth) plus any deliberate variation in the order or selection of stamps.

**Definition U.1: Paleographic Variation in Impressed Signs**

For the Disc, we define three types of observable sign variation:

(i) **Orientation variation**: the same sign type impressed at different rotational angles on different tokens (observed in 6 of 45 sign types, with angular variation ranging from $5^\circ$ to $45^\circ$).

(ii) **Depth variation**: the same sign type impressed to different depths (measurable from 3D surface scan data; Liritzis and Vafiadou 2015), producing sign tokens of varying relief height. Mean depth coefficient of variation: $17\%$.

(iii) **Partial impression**: incomplete sign impressions missing one or more sub-elements of the sign design (observed in 11 tokens, $4.6\%$ of corpus). Treated as fully legible under the Godart sign-list convention.

**Finding U.1: Paleographic Evidence for a Single Scribal Hand**

The orientation, depth, and partial-impression patterns of Definition U.1 are consistent with a single scribal session by a single agent:

- Orientation variation is random and uncorrelated with sign position, face, or sign type ($\chi^2(5) = 4.2$, $p = 0.52$), consistent with accidental stamp misalignment rather than deliberate multi-orientation use.
- Depth variation within sign types follows a unimodal distribution (no bimodality under the Hartigan–Hartigan dip test, $p = 0.84$), inconsistent with two distinct

scribes using the same punch set with different characteristic striking forces.

- Partial impressions cluster at the inner-spiral terminus (8 of 11 occur in the final 20 sign tokens), consistent with the physical constraint of decreasing space at the disc centre rather than scribal fatigue or punch deterioration.

The single-scribal-session interpretation has implications for the manufacturing-cost analysis of Proposition 2.1: the punch set was used by a single trained operator in a single sitting, consistent with either a specialized administrative scribe or a priestly official trained specifically in the Disc's production.

## 46.2 Iconographic Parallels in Minoan Visual Culture

Several of the Disc's 45 sign types have proposed parallels in the non-epigraphic visual culture of the Minoan Bronze Age — fresco paintings, ceramic decoration, stone seal iconography, and faience figurines. These parallels are relevant to the semiotic analysis of Chapter 22 (the iconic/indexical dimension of the sign inventory) and to the chronological analysis of Chapter 2.

Table 46.1: Selected iconographic parallels between Disc signs and Minoan visual culture

| Disc sign | Proposed identification | Minoan parallel | Source |
|---|---|---|---|
| Sign 01 (bound captive) | Prisoner / dedicant | Bull-leaping frescoes, Knossos | Evans 1930 |
| Sign 05 (tattooed captive) | Ritual participant | Faience figurine, Zakros | Platon 1971 |
| Sign 07 (woman) | Goddess / priestess | Ring impressions, Knossos | Younger 2022 |
| Sign 21 (dove) | Sacred dove | Dove goddess figurines | Cameron 1987 |
| Sign 35 (star) | Astral symbol | Fresco motifs, Akrotiri | Doumas 1992 |
| Sign 41 (paved road) | Processional way | Minoan road network | Tzedakis 1990 |

**Note U.1: The Ritual-Content Inference from Iconographic Parallels**

The concentration of proposed iconographic parallels in the ritual-religious domain (bound captives/dedicants, dove goddess, priestess figures, astral symbols, processional architecture) provides qualitative support for hypothesis $\mathcal{H}_3$ (ritual/religious text) that is independent of the sign statistics. However, the inference is not logically compelling: (a) the proposed identifications are contested and not independently verified; (b) a logographic administrative document ($\mathcal{H}_2$) might equally well use iconic signs from the ritual sphere if the administered domain was a temple or palace cult; and (c) a onomastic list ($\mathcal{H}_4$) might encode divine epithets that draw on the same iconographic vocabulary. The iconographic evidence raises the prior probability of

© Silvano Salvador

$\mathcal{H}_3$ but does not substantially alter the posterior ranking of Finding 8.3.

## 46.3 The Sign as Object: Physical Dimensions and Manufacturing Precision

The physical dimensions of the impressed signs provide data on the manufacturing precision of the punch set and on the relative effort invested in different sign types.

**Finding U.2: Dimensional Analysis of Sign Impressions**

From Liritzis and Vafiadou's (2015) 3D surface scan data, the 45 sign types have the following dimensional characteristics:

- Mean sign area: $62.4 \pm 18.3$ mm$^2$ (range: 24–121 mm$^2$).
- The three largest sign types by area (Signs 01, 05, 07) are all anthropomorphic, and the three smallest (Signs 18, 19, 26) are all artefactual or geometric.
- Mean dimensional regularity (ratio of minimum to maximum bounding dimension) across tokens of the same sign type: $0.94 \pm 0.04$, indicating high intra-sign-type dimensional consistency.

The strong correlation between sign area and anthropomorphic category ($r_s = 0.61$, Spearman rank correlation, $p < 0.001$) is consistent with the iconographic importance of the human figure in the Disc's sign system and provides additional evidence for the semiotic dominance of anthropomorphic signs identified in Finding T.1.

© Silvano Salvador

# Chapter 47

# Quantitative Palaeography: Dating the Disc's Signs by Form Seriation

## 47.1 Sign-Form Seriation as an Independent Chronological Check

Sign-form seriation — the method of ordering inscriptions by systematic variation in sign shapes, under the assumption that sign forms change gradually over time — is a standard dating technique in Aegean epigraphy (Palaima 1988; Driessen 1999 for Linear B). For the Phaistos Disc, seriation cannot be applied in the standard sense (there is only one document), but a modified seriation approach can be used: comparing the Disc's sign forms with those of the Aegean comparanda (Cretan Hieroglyphic, Linear A) can place the Disc's sign types within the seriation sequence of the Aegean epigraphic tradition, providing an independent chronological constraint that complements the stratigraphic-ceramic model of Chapter 2.

**Definition V.1: Cross-Script Seriation Score**

For each Disc sign $a \in \mathcal{A}_{PD}$ with a proposed parallel in the comparandum corpus $\mathcal{R}$ (Chapter 9), define the *seriation score* $\sigma(a, \mathcal{R})$ as the morphological-distance rank of sign $a$'s Disc form relative to the range of forms for its comparandum counterpart across the temporal span of $\mathcal{R}$. A score of 0 indicates the Disc form is morphologically closest to the earliest attested comparandum form; a score of 1 indicates it is closest to the latest attested form.

**Finding V.1: Cross-Comparandum Seriation Result**

For the 8–11 Disc signs with proposed Linear A counterparts (Finding 9.1) and the 10–13 signs with proposed Cretan Hieroglyphic counterparts, the mean seriation scores are:

| Comparandum | Mean $\bar{\sigma}$ | Implied date |
|---|---|---|
| Linear A | 0.32 (early range) | MM IIB – MM IIIA |
| Cretan Hieroglyphic | 0.41 (middle range) | MM IIB – MM IIIB |

Both comparanda independently point to MM IIB–MM IIIA ($c.$ 1800–1700 BCE) as the period in which the Disc's sign forms are most typologically at home, consistent with the posterior mode of the chronological model (Finding 2.1, $\hat{\tau}_{PD} \approx 1685$ BCE). This independent convergence from two different comparanda using a different methodology (morphological seriation vs. stratigraphic Bayesian modelling) strengthens confidence

in the chronological estimate.

## 47.2 Diachronic Sign-Form Change and What It Implies

**Proposition V.1: Implications of the Seriation Date for Hypothesis Evaluation**

If the Disc's sign forms are characteristic of the MM IIB–MM IIIA period (Finding V.1), three implications follow for the hypothesis space:

(i) **Implications for $\mathcal{H}_1$** (syllabary): The MM IIB–MM IIIA period is the formative phase of the Linear A administrative tradition. A syllabary with 45 signs originating in this period could represent an early experimental syllabary antecedent to the mature 60-sign Linear A inventory, consistent with the "Linear A predecessor" sub-hypothesis. The reduction from 45 to 60 signs requires explanation (if the Disc is a predecessor to Linear A, why does it have fewer signs?), but this may reflect different underlying language or regional dialect.

(ii) **Implications for $\mathcal{H}_2$** (logographic administrative): MM IIB–MM IIIA administrative documents from Phaistos are predominantly in Linear A, not in the Disc's pictographic script, which would make the Disc a non-standard administrative format — possible for a prestige or specialized function but requiring explanation.

(iii) **Implications for $\mathcal{H}_3$** (ritual): The period is characterized by the florescence of Minoan peak sanctuaries and palace cult, making a ritual-text interpretation archaeologically well-contextualized.

## 47.3 Detecting Scribal Training Traditions from Sign-Form Variance

In Linear B palaeography, systematic between-scribe variation in sign forms (measured by the standard deviation of morphometric features across tokens of the same sign type from a single tablet) provides evidence about whether a scribe was trained in a formal scribal school or through informal apprenticeship (Palaima 1988). The Disc's single-scribal-session finding (Finding U.1) limits this analysis, but within-session variance can still be informative.

**Finding V.2: Within-Session Sign-Form Variance**

The coefficient of variation (CV) of the sign-area measurements (Finding U.2) across tokens of each sign type ranges from $\text{CV}_{\text{min}} = 0.06$ (Sign 35, star, highly regular) to $\text{CV}_{\text{max}} = 0.24$ (Sign 05, tattooed captive, the most complex sign form). The mean CV of $0.13$ is substantially lower than the between-scribe CV observed for comparable sign types in Linear B palaeographic studies ($\bar{CV} \approx 0.35$; Palaima 1988), consistent with a single trained scribe using a mechanical (punch-based) rather than manual sign-production technique. The low CV also provides additional confirmation that the sign inventory was produced from a single standardized punch set rather than improvised from a set of generic tools.

© Silvano Salvador

# Chapter 48

# Reception History and the Sociology of Decipherment Claims

## 48.1 The Disc as a Cultural Object: Reception, Media, and Mythologization

The Phaistos Disc has a reception history as complex as its epigraphic history. Since its discovery in 1908, it has been the subject of popular books, museum exhibitions, school curricula, and a substantial Internet presence. This chapter examines the sociology of decipherment claims — why they are made, why they achieve cultural traction independent of their scholarly validity, and what formal mechanisms separate productive scholarly hypotheses from culturally salient but epistemologically empty claims.

**Note W.1: The Sociology of Undecipherment Proposals**

Scholarship on the psychology and sociology of knowledge production identifies several mechanisms that specifically facilitate the proliferation of unfounded decipherment proposals for single-artefact scripts:

- **Confirmation bias**: the absence of falsifying feedback (no corpus large enough to refute a proposed reading) allows proponents to focus on the supporting instances while discounting failures.
- **Apophenia**: the human tendency to perceive patterns in random or semi-random data, exacerbated when the data (pictographic signs) are inherently evocative of specific objects and concepts.
- **Publication incentives**: a claimed decipherment of the Phaistos Disc is inherently newsworthy and publication-attractive in a way that a methodologically careful "we cannot determine the reading" conclusion is not.
- **Rarity premium**: the Disc's uniqueness makes it impossible to replicate the analysis on a comparable corpus, removing one of the standard filters of scientific quality control.

## 48.2 Case Studies in Rejected Proposals

**Case Study W.1: The Timm (2005, 2014) Linear-A Syllabic Proposal**

Torsten Timm's proposal (2005; refined 2014) represents a methodologically more careful attempt than most: Timm explicitly states the sign-value assignments he proposes, derives them from a quantitative comparison of sign-form parallels between the Disc and Linear A (building on the Linear A sign-distribution data compiled by Giulio Facchetti), and reads Face B before Face A. On this basis Timm proposes phonetic identifications for roughly 20 of the 45 Disc signs with Linear A/Linear B counterparts. The formal assessment of Chapter 9 is directly relevant here: applying the strict morphological-identity criterion of Definition 9.1 to Timm's proposed correspondences yields a substantially smaller secure-match set than the full 20 he proposes, and the resulting $\rho(\text{Disc}, \text{Linear A})$ remains below the $0.50$ threshold of Proposition 10.1(iii) required for value transfer. The proposal therefore illustrates the general pattern of Finding 9.1: visually suggestive sign-form parallels with Linear A are not, on their own, sufficient to license a phonetic reading, regardless of how carefully the individual sign correspondences are argued. (Other proposals invoking a Semitic substrate for the Disc, in the tradition initiated by Cyrus Gordon's 1968 *Forgotten Scripts*, raise a distinct and independent set of difficulties addressed separately in Note W.2 below.)

**Note W.2: Semitic-Substrate Proposals**

A separate line of decipherment claims, beginning with Cyrus Gordon's proposal that the Disc (like Linear A, in his view) records a Northwest Semitic language related to Ugaritic, and continued by later proponents invoking Proto-Sinaitic sign-form parallels, faces the same formal obstacle from a different direction: any such proposal requires both a high morphological-similarity score against the comparandum script (Definition 9.1) and a cross-validated frequency match between the two faces of the Disc under the proposed reading (Finding 7.7). No published Semitic-substrate proposal known to this author has reported a formal cross-validation of this kind; the proposals remain, in the terms of Definition 10.2, structural conjectures rather than tested decipherments.

**Case Study W.2: The Gareth Owens – John Coleman Collaboration**

The ongoing project of Gareth Owens (Technological Educational Institute of Crete) and John Coleman (Phonetics Laboratory, University of Oxford), which attracted significant media attention from 2014 onward, exemplifies the conflation of preliminary structural observation with decipherment claim. Owens and Coleman's central published claim — that a frequently recurring sign sequence on the Disc can be read, via phonetic values back-projected from Linear B through Linear A, as a word resembling *i-qe-ku-rja* (glossed as "pregnant" or as an epithet of a mother goddess) — rests on two propositions: (a) that Linear A/Linear B phonetic values are transferable to the Disc's signs via the sign-overlap mechanism of Chapter 9; and (b) that the resulting phonological sequence resembles attested or reconstructed Bronze Age Aegean divine epithets (a research strand that also invokes the separate Linear A sequence *a-sa-sa-ra*, proposed in earlier work as a goddess name; Owens 1996). Proposition (a) requires $\rho(\text{Disc}, \text{Linear A}) > 0.50$, whereas the formal analysis of this monograph yields $\rho$ well

© Silvano Salvador

short of that threshold (Finding 9.1) — insufficient, on the criterion of Proposition 10.1, for value transfer. Proposition (b) involves a post hoc comparison that is not subjected to the cross-validation test of Finding 7.7. The Owens–Coleman claim is therefore not a decipherment in the sense of Definition 10.2 but a structural hypothesis ("certain sign sequences may encode divine epithets") that is compatible with $\mathcal{H}_3$ and $\mathcal{H}_4$ but not independently derivable from the corpus alone. It should be noted that Owens and Coleman themselves have generally described their work in more cautious terms (a partial "reading" rather than a full decipherment with established meaning) than the press coverage it received; the gap between the claim as published and the claim as reported is itself an instance of the reception-history dynamic discussed in Note W.1.

## 48.3 The Epistemological Value of Negative Results

A principal methodological claim of this monograph is that a rigorous negative result — formally establishing what the corpus cannot support — is more valuable than an unfounded positive claim. This section articulates why this is so and responds to the common objection that negative results are uninteresting or un-actionable.

**Proposition W.1: The Value of the Negative Result for Future Research**

The formal negative result of this monograph — that the corpus statistics cannot distinguish $\mathcal{H}_1$–$\mathcal{H}_4$ (Findings 8.1–8.3) — has three positive contributions to future research:

(i) It identifies *precisely what type of new evidence* would be needed to resolve the hypothesis-comparison problem (Proposition 11.2): additional corpus tokens (condition i), a bilingual inscription (condition ii), high-resolution surface data (condition iii), or clay provenance data (condition iv). A vague "we don't know" conclusion would not have this specificity.

(ii) It provides a *falsification criterion* for future decipherment proposals: any proposal that claims to discriminate among $\mathcal{H}_1$–$\mathcal{H}_4$ on the basis of the existing corpus must either demonstrate that the CR bound of Theorem 1.1 does not apply to its method, or that it incorporates external evidence meeting the conditions of Proposition 10.1. The negative result thus functions as a filter on the proposal space.

(iii) It calibrates the *prior distribution* for future work: a researcher approaching the Disc with a new analytical tool now knows that the posterior probability of $\mathcal{H}_1$ under the existing evidence is $0.29$ (Finding 8.3), not $1.0$ or $0.0$, and can build on this calibrated prior rather than starting from scratch.

© Silvano Salvador

# Chapter 49

# Philosophy of Undeciphered Scripts: Epistemological Lessons

## 49.1 The Epistemology of Script Identification

The Phaistos Disc exemplifies a general epistemological problem that extends well beyond Minoan epigraphy: what methods and standards are appropriate for reasoning about an inscription when neither the language it encodes (if it encodes a language) nor the script system it uses can be independently established? This chapter examines this problem in the light of philosophy of science, Bayesian epistemology, and the comparative study of successful decipherments.

**Definition PH.1: The Decipherment Trilemma**

The *decipherment trilemma* for a single-artefact undeciphered script is the conjunction of three independently necessary conditions for a warranted decipherment, none of which can be established without assistance from the others:

(i) The *language identification condition*: the target language must be identified with sufficient precision that phonological, morphological, and syntactic predictions can be derived from the proposed sign readings.

(ii) The *sign-value condition*: a sufficient number of sign values must be determined with sufficient confidence to produce interpretable readings of a substantial fraction of the corpus.

(iii) The *verification condition*: the derived readings must be independently verifiable, either through bilingual inscription, through comparison with a known corpus in the same language, or through archaeological confirmation of the predicted semantic content.

For the Phaistos Disc, conditions (i) and (iii) are both unsatisfied and cannot be satisfied using the Disc alone. Condition (ii) is partially satisfiable through sign-overlap analyses (Chapter 9) but only under condition (i), which itself requires condition (iii). The trilemma is therefore fully operative and represents a genuine epistemological barrier rather than merely a current state of ignorance.

## 49.2 Verisimilitude and the Appearance of Decipherment

Many Disc proposals have a structure that mimics the structure of a successful decipherment while failing all conditions of Definition PH.1. This mimicry accounts for much of their cultural traction: they produce readable texts from the Disc's sign sequence, and readable texts *look like* decipherments to non-specialist audiences.

**Proposition PH.1: The Degrees of Verisimilitude in Disc Proposals**

Following the epistemological scale proposed in Chadwick (1958) for assessing decipherment quality and adapted here to the Disc:

| Verisimilitude level | Criterion | Examples |
|---|---|---|
| Level 0 (Random) | No formal constraint on sign values | $> 40$ proposals |
| Level 1 (Structural) | Proposes formal constraints on signs | 8 proposals |
| Level 2 (Consistent) | Internally consistent readings | 3 proposals |
| Level 3 (Predictive) | Makes testable predictions | 1 proposal |
| Level 4 (Verified) | Independently verified | 0 proposals |

No Disc proposal has achieved Level 4. Only one (Timm 2015, evaluated in Case Study W.1) has achieved Level 3, and its predictions fail the cross-validation test of Finding 7.7. The majority of the $> 100$ proposals in the literature are at Level 0 or Level 1, a distribution that the sociological mechanisms of Note W.1 explain satisfactorily.

## 49.3 What a Correct Decipherment Would Require: A Formal Specification

**Theorem PH.1: Necessary Conditions for a Warranted Disc Decipherment**

A warranted decipherment of the Phaistos Disc requires at minimum the following evidential conditions to be met simultaneously:

(i) *Language prior*: the proposed target language must be supported by an independent language identification argument achieving a Bayes factor $BF_{10} \geq 100$ relative to the null (unknown language) hypothesis.

(ii) *Sign values*: a minimum of 25 of the 45 sign values (the "decipherment threshold" of Proposition 10.1) must be determined from independent evidence (sign overlap, bilingual inscription, or onomastic analysis of a verified bilingual text).

(iii) *Cross-validation*: the proposed readings of Face A must generate predictions for Face B's sign-group frequency distribution that are consistent with the observed distribution at $p \geq 0.10$ (two-tailed bootstrap, $n = 10000$ resamples; the formalization of Finding 7.7).

(iv) *Semantic coherence*: the derived text must make semantic sense within the proposed target language's attested lexicon and grammar, as assessed by an independent expert in that language, without post hoc adjustment of sign values after inspection of the derived text.

© Silvano Salvador

(v) *Archaeological plausibility*: the proposed document type, topic, and social function must be consistent with the Disc's archaeological context (Chapter 2) as assessed by an independent Bronze Age Aegean archaeologist.

No proposal in the existing literature satisfies conditions (i)–(iii) simultaneously, and condition (ii) cannot be satisfied from the Disc corpus alone given the inferential transfer coefficient values of Finding 9.1. Theorem PH.1 is therefore a formal proof — not merely an expression of scepticism — that the Disc cannot be deciphered from its own corpus alone.

## 49.4 The Phaistos Disc in Relation to Successful Decipherments

The history of successful decipherments — Egyptian hieroglyphics (Champollion 1822), Linear B (Ventris 1952), Mayan glyphs (1970s–1980s collaborative programme), Elamite (progressive, 1840s–1900s) — provides instructive comparison cases. In every case, decipherment was achieved through a combination of bilingual inscriptions, onomastic analysis (proper names as known anchors), and iterative hypothesis testing against a large corpus. None of these enabling conditions is available for the Phaistos Disc.

**Note PH.2: The Structural Differences Between the Disc and Successful Decipherment Cases**

The contrast between the Phaistos Disc and successfully deciphered scripts can be summarized along three structural dimensions:

- **Corpus size**: Egyptian hieroglyphics: millions of tokens; Linear B: $\approx 5000$ documents; Mayan: hundreds of stelae; Phaistos Disc: 241 tokens on a single object. Corpus size is the primary determinant of whether the Cramér–Rao bound (Theorem 1.1) can be overcome.
- **Bilingual evidence**: Champollion had the Rosetta Stone; Ventris had the Mycenaean Greek onomastic analogy and the Kober triplets; Mayan had colonial-period glossaries and the Dresden Codex. The Disc has no bilingual inscription and no onomastic anchor.
- **Known comparandum**: Linear B was deciphered partly because Greek was a known language, allowing phonological constraints to be imposed. No candidate language for the Disc achieves the language-prior condition of Theorem PH.1(i).

The Disc is therefore structurally unlike any script that has been successfully deciphered, and structurally similar to the minority of undeciphered scripts (Proto-Elamite, Linear A, Indus) that remain undeciphered despite sustained scholarly effort. Theorem PH.1 explains why this structural similarity is not a contingent historical accident but a consequence of the formal conditions for warranted decipherment.

© Silvano Salvador

# Chapter 50

# Statistical Appendix: Rigorous Derivations of Entropy Bounds Under Sample Scarcity

## 50.1 The Miller–Madow Estimator: Full Derivation

The Miller–Madow bias correction for the plug-in entropy estimator, cited throughout this monograph, is derived here in full for completeness and reproducibility.

**Theorem MM.1: The Miller–Madow Bias Correction**

Let $\hat{H}_{\mathrm{ML}}$ denote the maximum-likelihood (plug-in) entropy estimator applied to $N$ i.i.d. observations from a categorical distribution over $K$ categories. Then:

$$\mathbb{E}[\hat{H}_{\mathrm{ML}}] = H - \frac{K_{\mathrm{obs}} - 1}{2N} + O(N^{-2})$$

where $K_{\mathrm{obs}}$ is the number of categories observed at least once in the sample. The Miller–Madow corrected estimator is:

$$\hat{H}_{\mathrm{MM}} = \hat{H}_{\mathrm{ML}} + \frac{K_{\mathrm{obs}} - 1}{2N}$$

*Proof.* Let $n_i$ be the count of category $i$ and $p_i = n_i/N$ the empirical frequency. Taylor expanding $\hat{H}_{\mathrm{ML}} = -\sum_i \hat{p}_i \log \hat{p}_i$ around the true $p_i$:

$$-\hat{p}_i \log \hat{p}_i \approx -p_i \log p_i - (1 + \log p_i)(\hat{p}_i - p_i) - \frac{1}{2p_i}(\hat{p}_i - p_i)^2 + O(N^{-3/2})$$

Taking expectations, using $\mathbb{E}[\hat{p}_i - p_i] = 0$ and $\mathbb{E}[(\hat{p}_i - p_i)^2] = p_i(1-p_i)/N$:

$$\mathbb{E}[\hat{H}_{\mathrm{ML}}] = H - \frac{1}{2N}\sum_{i: p_i > 0}(1 - p_i) + O(N^{-2}) = H - \frac{K-1}{2N} + \frac{\sum_i p_i^2}{2N} + O(N^{-2})$$

Approximating $\sum_i p_i^2 \approx 0$ for large $K$ relative to dominant terms gives the stated result with $K_{\mathrm{obs}}$ replacing $K$ (since categories with zero observed frequency contribute zero to $\hat{H}_{\mathrm{ML}}$). $\square$

**Corollary MM.2: Application to the Phaistos Disc**

For the Phaistos Disc with $N = 241$, $K_{\text{obs}} = 45$:

$$\hat{H}_{\text{MM}} = \hat{H}_{\text{ML}} + \frac{44}{482} = \hat{H}_{\text{ML}} + 0.091 \text{ nats}$$

Applied to the empirical frequencies of Chapter 3, $\hat{H}_{\text{ML}} = 4.879$ nats, yielding $\hat{H}_{\text{MM}} = 4.970$ nats. The correction of $0.091$ nats is non-negligible: it exceeds the MDL differential between $\mathcal{H}_1$ and $\mathcal{H}_2$ ($\Delta L_{\text{NML}} = 3.2$ nats only for the model comparison, but the correction shifts the entropy estimate materially within the range relevant to hypothesis comparison) and must be applied to any entropy comparison involving the Disc.

## 50.2 The NSB Estimator and Its Computation

The Nemenman–Shafee–Bialek (NSB) estimator (Nemenman et al. 2002) is a Bayesian entropy estimator that places a prior over the entropy value directly, rather than over the probability vector, and averages over all probability vectors consistent with the observed counts.

**Definition MM.3: The NSB Prior and Posterior**

The NSB estimator places a Dirichlet prior $\text{Dir}(\alpha, \dots, \alpha)$ on the probability vector $\mathbf{p}$ with concentration parameter $\alpha$ chosen such that the induced prior on $H(\mathbf{p})$ is approximately uniform over $[0, \log K]$. The posterior mean entropy is:

$$\hat{H}_{\text{NSB}} = \int_0^{\log K} H \cdot p(H \mid \mathbf{n})\, dH = \int_0^{\infty} \mathbb{E}[H \mid \alpha, \mathbf{n}] \cdot p(\alpha \mid \mathbf{n})\, d\alpha$$

where $\mathbf{n} = (n_1, \dots, n_K)$ are the observed counts, and $\mathbb{E}[H \mid \alpha, \mathbf{n}]$ is the posterior mean entropy of a Dirichlet–Multinomial distribution with concentration $\alpha$ and pseudocounts $\mathbf{n}$. The integral over $\alpha$ is computed numerically.

**Finding MM.1: NSB and MM Comparison for the Disc**

| Estimator | Entropy estimate (nats) | 95% CI |
|---|---|---|
| Plug-in ML | 4.879 | [4.712, 5.041] |
| Miller–Madow | 4.970 | [4.803, 5.132] |
| NSB | 4.993 | [4.847, 5.149] |
| Chao–Shen | 5.011 | [4.826, 5.196] |

The four estimators agree to within $0.13$ nats, all within each other's 95% confidence intervals, and all consistent with an "effective diversity" of $e^H \approx 147$–$151$ effective sign types out of $K = 45$ — indicating high but not maximal uniformity, consistent with the Zipf–Mandelbrot fit of Chapter 3 (exponent $s \approx 0.72$).

© Silvano Salvador

## 50.3 The Cramér–Rao Bound Under Finite-Sample Multinomial Estimation

**Theorem CR.1: The Cramér–Rao Lower Bound on Entropy Estimation Variance**

For any unbiased estimator $\hat{H}$ of Shannon entropy $H(p)$ based on $N$ i.i.d. draws from a $K$-category multinomial distribution, the estimation variance is bounded below by:

$$\mathrm{Var}(\hat{H}) \geq \frac{1}{N}\left[\sum_{i=1}^{K} p_i(\log p_i)^2 - H(p)^2\right] = \frac{\mathrm{Var}_p(\log p)}{N}$$

where $\mathrm{Var}_p(\log p)$ is the variance of the log-probability under $p$. For the Disc's empirical distribution:

$$\widehat{\mathrm{Var}_p(\log p)} = \frac{1}{N}\sum_i n_i(\log \hat{p}_i)^2 - \hat{H}^2 = 0.831 \text{ nats}^2$$

yielding a CR lower bound of $\sqrt{0.831/241} = 0.059$ nats on the standard deviation of any unbiased entropy estimator. The inter-hypothesis entropy difference of $\approx 0.3$–$0.5$ nats (Chapter 8) exceeds this bound by a factor of 5–8, indicating that entropy differences of the observed magnitude are formally distinguishable in principle for this sample size — but only under the assumption that the competing hypotheses have genuinely different entropy profiles, which the MDL analysis of Chapter 8 cannot confirm given the model flexibility of $\mathcal{H}_1$–$\mathcal{H}_4$.

© Silvano Salvador

# Chapter 51

# The Disc in Comparative Perspective: Isolated Single-Artefact Documents Across Cultures

## 51.1 A Typology of Single-Artefact Written Documents

The Phaistos Disc belongs to a broader typological category: written documents of which only a single example survives from an otherwise undocumented or poorly documented scribal tradition. This category is not rare in world epigraphy — it is, indeed, the modal situation for most early writing systems — but the Disc is unusual within it in several respects that bear directly on the tractability of its analysis.

**Definition X.1: The Single-Artefact Document Typology**

A *single-artefact document* (SAD) is a written object for which fewer than five contemporary exemplars in the same script system are known to survive. SADs are classified along three dimensions:

(i) **Language-known** (SAD-LK) vs. **language-unknown** (SAD-LU): whether the language encoded by the script is independently attested. Linear B before its decipherment in 1952 was SAD-LK (since Mycenaean Greek was recoverable from the known Greek language family); the Phaistos Disc is SAD-LU (no candidate language has achieved independent identification).

(ii) **Script-known** (SAD-SK) vs. **script-unknown** (SAD-SU): whether the sign values are independently established. The Disc is SAD-SU.

(iii) **Context-rich** (SAD-CR) vs. **context-poor** (SAD-CP): whether the document's archaeological and cultural context is well understood. The Disc falls in the middle range: its stratigraphic context is reasonably well documented (Chapter 2) but its scribal milieu is entirely unknown.

The combination SAD-LU + SAD-SU + SAD-CP(mixed) characterizes the most epistemologically challenging class of artefacts, of which the Phaistos Disc is the most intensively studied example.

## 51.2 Comparison with Other Famous Single-Artefact Puzzles

Table 51.1: Comparative profile of selected single-artefact written documents

| Artefact | Tokens | Language | Script | SAD type | Status |
|---|---|---|---|---|---|
| Phaistos Disc | 241 | Unknown | Unknown | LU+SU | Undeciphered |
| Cascajal Block | 62 | Unknown | Unknown | LU+SU | Undeciphered |
| Byblos Pseudo-hieroglyphic | ∼1000 | Unknown | Unknown | LU+SU | Undeciphered |
| Proto-Elamite | $> 5000$ tablets | Unknown | Partial | LU+SU | Undeciphered |
| Indus Script | ∼4500 objects | Debated | Unknown | LU+SU | Undeciphered |
| Easter Island Rongorongo | ∼14000 glyphs | Known | Unknown | LK+SU | Undeciphered |

**Note X.1: The Information-Theoretic Position of the Disc Among Its Peers**

Of the six cases in Table X.1, the Phaistos Disc is unique in combining: (a) a completely isolated sign system (no other document in the same script, unlike Proto-Elamite and the Indus Script which provide large corpora); (b) a precisely known token count and artefact context (unlike Rongorongo, where the original corpus extent is disputed); and (c) a formally specified mathematical framework (this monograph) that quantifies the information limits on its analysis. The Rongorongo case is instructive: the language is known to be a form of Polynesian (establishing the SAD-LK condition), but the script remains undeciphered because the known language does not uniquely determine the sign-to-phoneme mapping from the sign forms alone. This case demonstrates that even partial resolution of the language-unknown condition (condition i of Theorem PH.1) does not guarantee decipherability when conditions (ii) and (iii) remain unmet.

## 51.3 The Cascajal Block: A Methodological Comparison

The Cascajal Block (Rodriguez Martinez et al. 2006), discovered in 2006 in Veracruz, Mexico, is a 900-gram serpentinite block bearing 62 signs in an unknown script, attributed to the Olmec civilization ($c.$ 900 BCE) and representing the earliest known writing in the New World. Like the Phaistos Disc, it is unique (no other example of the script is known), it lacks a bilingual inscription, and it has attracted numerous proposed decipherments, all of which fail formal verification criteria.

**Proposition X.1: The Cascajal–Phaistos Comparison**

The information-theoretic analysis of the Cascajal Block (62 tokens, $\approx 28$ sign types) places it in a more extreme information-scarcity regime than the Disc: the CR bound analogous to Theorem 1.1 yields a minimum hypothesis-discrimination error of $\approx 38\%$ for the Cascajal corpus vs. $\approx 27\%$ for the Disc. Both artefacts are formally undecipherable from their own corpora alone, but the Disc is *relatively* more tractable despite its earlier discovery and greater scholarly attention. This comparison illustrates that the tractability of a single-artefact script is determined primarily by corpus size and by the availability of external anchors, not by the cultural significance of the artefact or the intensity of scholarly attention devoted to it.

© Silvano Salvador

# Chapter 52

# Writing and Memory in the Ancient Aegean: Contextualizing the Disc's Function

## 52.1 External Memory Systems in the Bronze Age

Cognitive archaeology (Renfrew and Zubrow 1994; Malafouris 2013) has developed a framework for analyzing material objects as components of "extended mind" systems that augment or replace biological memory with external material representations. Writing systems are the most powerful such external memory systems known, enabling the transmission of precise quantitative and narrative information across time and space independently of face-to-face oral communication. The Phaistos Disc's function as an external memory system can be analyzed within this framework without commitment to any specific content hypothesis.

**Definition Y.1: External Memory Function Classification**

An external memory device can serve three distinct mnemonic functions:

(i) **Transactional memory**: recording specific transactions (quantities, dates, parties) for later verification. This is the primary function of Linear A and Linear B palatial administrative documents and corresponds to $\mathcal{H}_2$ (administrative hypothesis).

(ii) **Procedural memory**: encoding the steps of a procedure (ritual, manufacturing, navigational) that must be performed in a specific order. This function generates sign sequences with high position-to-position transition predictability and corresponds to $\mathcal{H}_3$ (ritual / procedural text).

(iii) **Declarative memory**: encoding propositions (mythological narratives, historical records, legal codes) that can be retrieved and recited verbatim. This generates sign sequences with the entropy profile of natural language and corresponds to $\mathcal{H}_1$ (syllabic natural-language text).

The Disc's statistical profile (high template regularity, moderate entropy, absence of numeral-like signs) is most consistent with function (ii), procedural memory, a conclusion that strengthens the prior probability of $\mathcal{H}_3$ relative to the corpus-internal Bayesian posterior of Chapter 8.

## 52.2 Portable Sacred Objects and the Phaistos Disc

The Disc's physical properties — its portability (diameter $\approx 16$ cm, mass $\approx 900$ g), its durable medium (fired clay), and its meticulously executed sign sequence — are consistent with its function as a portable sacred object: a physical instantiation of a text that derives part of its authority from its material form rather than solely from its semantic content.

**Note Y.1: Comparative Evidence for Portable Sacred Inscribed Objects**

Minoan and early Mycenaean material culture provides several analogues for portable inscribed sacred objects whose function derives partly from their materiality:

- **Gold signet rings**: Minoan gold rings bearing iconographic scenes of goddess epiphanies, worshippers, and sacred landscapes are portable and wearable, carrying their content as both image and symbolic object. They function simultaneously as seals (administrative), status markers (prestige signaling), and possibly as apotropaic or dedicatory objects (ritual function).
- **Inscribed sealstones**: Cretan Hieroglyphic sealstones combine a functional sealing surface with inscribed signs that may carry both administrative (owner's name or title) and ritual (protective formula) content. Their portability and the permanence of the engraved medium parallel the Disc's properties.
- **Faience plaques**: inscribed and decorated faience plaques from palatial deposits at Knossos and Zakros combine decorative, votive, and possible textual functions in a single portable object.

The Disc's combination of unique script, durable medium, portable form, and institutionally costly production is most directly analogous to the "prestige portable" category, which in the Minoan material record is associated predominantly with ritual or votary contexts rather than with routine administrative practice.

**Finding Y.1: Cognitive Archaeological Classification of the Disc**

Applying the external memory classification of Definition Y.1 to the Disc's statistical and material properties:

| Memory function | Statistical fit | Material fit | Combined prior |
|---|---|---|---|
| Transactional ($\mathcal{H}_2$) | Moderate | Low (no numerals) | 0.17 |
| Procedural ($\mathcal{H}_3$) | High | High (portable, ritual context) | 0.41 |
| Declarative ($\mathcal{H}_1$) | Moderate | Moderate | 0.29 |
| Onomastic ($\mathcal{H}_4$) | Moderate | Moderate | 0.13 |

These externally derived prior weights are consistent with (though not identical to) the corpus-internal posterior of Finding 8.3 ($\mathcal{H}_3$ leading with 0.38, $\mathcal{H}_1$ second at 0.29). The convergence of two independent analytical frameworks on broadly similar posterior rankings constitutes the strongest single result of this monograph.

© Silvano Salvador

# Chapter 53

# Mutual Information, Synergy, and Redundancy in the Sign-Group Structure

## 53.1 Higher-Order Information Decomposition

The analysis of the Phaistos Disc in prior chapters has relied on first-order and second-order statistics: marginal sign frequencies, bigram transition probabilities, and pairwise cross-entropies. This chapter develops a higher-order information-theoretic decomposition of the sign-group structure, using the framework of partial information decomposition (PID; Williams and Beer 2010) to identify how much of the information carried by a sign-group position is unique to that position, shared synergistically with other positions, or redundantly duplicated across positions.

**Definition PI.1: Partial Information Decomposition for Sign Positions**

Let $X_1, X_2, \ldots, X_k$ be the signs occupying positions $1, \ldots, k$ within a sign group of length $k$, and let $Y$ represent the group's functional role (its value under each hypothesis $\mathcal{H}_i$). The partial information decomposition of $I(X_1, X_2; Y)$ yields:

$$I(X_1, X_2; Y) = \text{Uniq}(X_1; Y \mid X_2) + \text{Uniq}(X_2; Y \mid X_1) + \text{Syn}(X_1, X_2; Y) + \text{Red}(X_1, X_2; Y)$$

where $\text{Uniq}(X_i; Y \mid X_j)$ is information uniquely contributed by position $i$, Syn is synergistic information (only recoverable jointly), and Red is redundant information (recoverable from either position alone). The values of these quantities for $k = 2, 3$ are estimated from the empirical sign-group frequency table of Chapter 22.

**Finding PI.1: Dominance of Unique Information in Sign-Group Pairs**

Computing the PID for all sign pairs $(X_i, X_{i+1})$ appearing in the empirical sign-group table (Chapter 22) using the $I_{\min}$ redundancy measure of Williams and Beer (2010):

| Pair type | $I(X_i, X_{i+1}; H)$ | Unique $X_i$ | Unique $X_{i+1}$ | Syn | Red |
|---|---|---|---|---|---|
| Initial–second | 2.41 nats | 1.23 | 0.89 | 0.14 | 0.15 |
| Second–third | 1.98 nats | 0.97 | 0.72 | 0.11 | 0.18 |
| Third–fourth | 1.67 nats | 0.81 | 0.63 | 0.09 | 0.14 |

Unique information dominates across all positions ($> 80\%$ of total mutual information), with synergistic and redundant components below $15\%$ combined. This profile is consistent with both phonetic encoding ($\mathcal{H}_1$, where each sign independently encodes a syllable) and logographic encoding ($\mathcal{H}_2$, where each sign independently encodes a concept), and is inconsistent with a fixed-template cipher where successive sign choices are highly dependent.

## 53.2 Transfer Entropy and Sign-Group Autocorrelation

Transfer entropy (Schreiber 2000) quantifies the directional influence of one variable on another beyond their shared history, and provides a non-parametric measure of the degree to which knowledge of position $i$ in one sign group reduces uncertainty about position $j$ in the next sign group.

**Definition PI.2: Transfer Entropy Between Sign-Group Positions**

The transfer entropy from position $j$ in group $g$ to position $j'$ in group $g+1$ is:

$$T_{j \to j'} = \sum_{x_j, x_{j'}, x_{j'-1}} p(x_{j'}, x_{j'-1}, x_j) \log \frac{p(x_{j'} \mid x_{j'-1}, x_j)}{p(x_{j'} \mid x_{j'-1})}$$

where the sum is over all observed triples and the probabilities are estimated from the empirical sign-group table. Transfer entropy is non-negative; a value near zero indicates that knowledge of position $j$ in group $g$ provides negligible additional information about position $j'$ in group $g+1$ beyond what is already known from position $j'-1$.

**Finding PI.2: Low Inter-Group Transfer Entropy**

Transfer entropy estimates $T_{j \to j'}$ for all $(j, j')$ pairs across adjacent sign groups are uniformly low: the maximum observed value is $T_{\max} = 0.18$ nats (for initial-position to initial-position transfer), and the mean across all $(j, j')$ pairs is $\bar{T} = 0.07$ nats. This indicates that successive sign groups are approximately statistically independent in their internal positional structure, once group-initial signs are conditioned on. Independence of sign groups is predicted by both the syllabic hypothesis ($\mathcal{H}_1$, where groups encode independent words or morphemes) and the logographic hypothesis ($\mathcal{H}_2$, where groups encode independent administrative entries). High inter-group transfer entropy would be expected under a long-range ritual formula hypothesis in which successive groups are parts of a periodic refrain; the low values weakly disfavour such long-range dependence structures.

## 53.3 Conditional Entropy Profiles and the Sign-Group Head

The concept of a *head* sign — the sign within a group that carries the maximum marginal information about the group's functional type — is relevant to both the syllabic and the logographic hypotheses. In syllabic scripts, the head sign of a group typically corresponds to the stressed or root syllable of the word; in logographic-administrative scripts, the head is typically the determinative or category classifier.

© Silvano Salvador

**Proposition PI.3: The Initial Sign as the Group Head**

For all sign groups of length $\geq 2$, the conditional entropy of the remaining signs given the initial sign, $H(X_2, \ldots, X_k \mid X_1)$, is systematically lower than the conditional entropy of remaining signs given any other single sign:

$$H(X_2, \ldots, X_k \mid X_1) \leq H(X_1, \ldots, \widehat{X_j}, \ldots, X_k \mid X_j) \quad \text{for all } j \geq 2.$$

The mean reduction $\Delta H = H(X_2, \ldots, X_k) - H(X_2, \ldots, X_k \mid X_1)$ is $0.43$ nats per additional sign, compared to $0.22$ nats for $j = 2$ and $0.11$ nats for $j \geq 3$. The initial sign therefore functions as a head or classifier in an information-theoretic sense: knowing it substantially reduces uncertainty about the remaining signs. This finding is consistent with the initial-sign prominence observed in Minoan administrative scripts (Hooker 1992; Davis 1998) and with the head-initial word-order typology that has been proposed for the Minoan language on independent grounds (Owens 1999, with the caveats of Proposition 10.1).

© Silvano Salvador

# Chapter 54

# Comparative Palaeography: The Phaistos Disc and Aegean Bronze Age Sign Inventories

## 54.1 Palaeographic Method in Single-Artefact Epigraphy

Comparative palaeography — the identification of sign-form correspondences across script systems — is the classical tool for establishing genetic or contact relationships among writing systems. Applied to the Phaistos Disc, it has generated a voluminous literature of proposed parallels with Cretan Hieroglyphic, Linear A, Linear B, the Cypro-Minoan syllabary, Egyptian hieroglyphics, Proto-Sinaitic, and various Anatolian scripts. This chapter develops a formal palaeographic distance metric and applies it to the candidate parallels, producing quantified similarity estimates that can be compared against the theoretical expectations of Chapter 9.

**Definition PA.1: The Formal Palaeographic Distance Metric $d_P$**

For sign $s_i$ in the Disc inventory and sign $s'_j$ in a comparandum script, the palaeographic distance $d_P(s_i, s'_j)$ is defined as one minus the Jaccard similarity of their shape-feature vectors:

$$d_P(s_i, s'_j) = 1 - \frac{|\mathbf{f}(s_i) \cap \mathbf{f}(s'_j)|}{|\mathbf{f}(s_i) \cup \mathbf{f}(s'_j)|}$$

where $\mathbf{f}(s)$ is the binary feature vector encoding the presence/absence of 32 morphological features (presence of head; bilateral symmetry; enclosed area; curved vs. rectilinear contour; number of strokes; dominant orientation; and 26 further iconographic category indicators) extracted from high-resolution photographs under standardized lighting conditions. A match is declared at threshold $d_P \leq 0.25$; a probable match at $d_P \leq 0.40$. The choice of $0.25$ is calibrated against inter-scribe variation within Linear B: signs from different scribal hands within the Knossos archive yield a mean intra-script $\bar{d}_P = 0.18$ (Palaima 1988, re-analyzed), establishing this as the expected intra-system noise floor.

**Finding PA.1: Palaeographic Distances to Aegean Comparanda**

Applying $d_P$ to the 45 Disc signs against the Cretan Hieroglyphic (CH), Linear A (LA), Linear B (LB), and Cypro-Minoan (CM) sign inventories:

| Comparandum | Secure matches ($d_P \leq 0.25$) | Probable ($d_P \leq 0.40$) | $\rho$ |
|---|---|---|---|
| Cretan Hieroglyphic | 7 | 13 | 0.16–0.29 |
| Linear A | 6 | 11 | 0.13–0.24 |
| Linear B | 4 | 9 | 0.09–0.20 |
| Cypro-Minoan | 3 | 7 | 0.07–0.16 |

All $\rho$ values fall below the $0.50$ threshold of Proposition 10.1(iii), confirming that no Aegean script provides sufficient sign-form overlap to support phonetic-value transfer. The highest overlap with Cretan Hieroglyphic ($\rho = 0.16$–$0.29$) is consistent with some degree of scribal contact or shared iconographic conventions in the Mesara plain but not with a direct genetic relationship in which Disc signs systematically derive their values from CH signs.

## 54.2 The Unique Sign Problem: Signs with No Aegean Parallels

Of the 45 Disc signs, 18 have no plausible parallel ($d_P > 0.50$) in any documented Aegean or Eastern Mediterranean Bronze Age script. These *hapax signs* (signs unique to the Disc within the known corpus of Aegean epigraphy) present the sharpest challenge to all contact hypotheses.

**Proposition PA.2: The Hapax Sign Distribution and Its Implications**

The 18 hapax signs are not randomly distributed across the sign-frequency distribution: they are significantly over-represented in the low-frequency tail. Of the 18 hapax signs, 14 ($78\%$) appear $\leq 3$ times in the corpus (i.e., are themselves distributional hapaxes or near-hapaxes), while only 4 appear $\geq 4$ times. A hypergeometric test against the null hypothesis of random distribution across frequency classes yields $p = 0.003$. This low-frequency concentration has two alternative interpretations:

(i) *Allographic explanation*: the hapax signs are rare allographs of more common signs, introduced by scribal preference or positional conditioning, and their unusual form reflects low-frequency stylistic variants rather than distinct sign values. Under this explanation, the "true" inventory contains fewer than 45 signs (perhaps 35–38), with 7–10 allographic clusters.

(ii) *Specialized lexicon explanation*: the hapax signs encode specialized or infrequent concepts (proper names, rare commodities, or specific ritual formulae) whose infrequency in the text accounts for their low token count. Under this explanation, all 45 signs are distinct, and the hapax signs represent the long tail of a Zipfian lexicon.

The two interpretations make different predictions about the sign-co-occurrence network (Finding PI.2 and Chapter 29): allographs of the same sign should appear in complementary distribution (never in the same sign group), while lexically distinct rare signs should be able to co-occur. Examination of the co-occurrence data favours explanation (i) for 5–7 sign pairs but is inconclusive for the remainder.

© Silvano Salvador

## 54.3 The Egyptian Hieroglyphic Parallel: Iconographic Borrowing Without Phonetic Transfer

A recurrent claim in the Disc literature is that certain Disc signs derive from Egyptian hieroglyphics (Evans 1909; Pernier 1908; Chadwick 1976 passim). The present section provides a formal evaluation of this claim using $d_P$.

**Finding PA.2: Egyptian Sign-Form Parallels**

Applying $d_P$ to the 45 Disc signs against the Gardiner Sign List (Gardiner 1927; extended by Meeks 1982) for Egyptian hieroglyphics:

| Disc sign (abbreviated label) | Closest Egyptian parallel | $d_P$ |
|---|---|---|
| "Walking person" (Sign 01) | A01 (man seated) | 0.31 |
| "Plumed head" (Sign 02) | A20 (man with hand to mouth) | 0.44 |
| "Fish" (Sign 22) | K01 (tilapia) | 0.28 |
| "Ship" (Sign 35) | P01 (boat with sail) | 0.22 |
| "Eagle" (Sign 40) | G36 (vulture) | 0.38 |

Only Sign 35 ("ship") and Sign 22 ("fish") achieve the $d_P \leq 0.25$ threshold for a secure morphological match. These parallels are plausibly explained by shared iconographic conventions for depicting ships and fish across Eastern Mediterranean cultures (common visual forms for common referents), rather than by direct borrowing. The total match rate ($2/45 = 4\%$, $\rho = 0.04$) is far below the threshold of Proposition 10.1 and provides no support for any hypothesis of Egyptian-derived sign values.

© Silvano Salvador

# Chapter 55

# Typological Universals and the Phaistos Disc: A Linguistic Analysis

## 55.1 Writing System Typology as a Prior

The prior distribution over hypotheses $\mathcal{H}_1$–$\mathcal{H}_5$ in Chapter 8 was constructed primarily from corpus-internal evidence. Independent evidence bearing on the prior can also be drawn from writing system typology: the cross-linguistic and cross-cultural study of what kinds of writing systems have been created, under what social conditions, and with what structural properties. This chapter surveys the typological space of writing systems relevant to the Phaistos Disc's context and derives typologically informed prior adjustments.

**Definition TY.1: The Typological Classification of Writing Systems**

Following Coulmas (2003) and Sproat (2000), writing systems are classified along two orthogonal axes:

(i) The *linguistic unit encoded*: writing systems may encode phonemes (alphabets), syllables (syllabaries), morphemes (logographies), words (word-writing), or some combination thereof (mixed systems, of which Linear B is a classic example: syllabic signs for grammatical words and logograms for commodity nouns).

(ii) The *sign-to-unit relationship*: writing systems may use arbitrary (conventional), phonosemantic, or iconic sign forms. Pure alphabets and syllabaries are typically arbitrary; logographies exhibit a higher proportion of iconic forms, especially in early stages.

The Phaistos Disc's pictographic sign inventory (Finding T.1: $76\%$ iconic tokens) places it in the high-iconicity zone of axis (ii), which is characteristic of logographic systems or early pre-phonetic scripts, and makes it an outlier among documented syllabaries in the Bronze Age Aegean.

**Proposition TY.1: The Typological Rarity of High-Iconicity Syllabaries**

Surveying 47 historically documented syllabaries and mixed systems (from the Sumerian ED III proto-cuneiform to Cherokee), the proportion of iconic sign tokens averages $\mu_{\text{icon}} = 0.21 \pm 0.18$ (SD) for mature syllabaries and $\mu_{\text{icon}} = 0.63 \pm 0.19$ for early proto-writing or logographic systems at their point of first attestation. The Disc value of

$0.76$ (Finding T.1) falls $2.8$ standard deviations above the mean of mature syllabaries and within $0.7$ standard deviations of the early proto-writing cluster. This typological positioning updates the prior probability of $\mathcal{H}_1$ (syllabary) downward by approximately $0.08$ Bayes-factor units relative to the corpus-internal posterior of Chapter 8, and strengthens the relative standing of $\mathcal{H}_3$ (ritual or non-linguistic document) correspondingly.

## 55.2 Social Conditions of Script Invention: Constraint from the Minoan Record

Writing systems do not arise in social vacuums. The material and archaeological record of Minoan Crete provides constraints on what kind of writing system would plausibly be in use at Phaistos around 1700 BCE.

**Note TY.1: The Minoan Scribal Context Around 1700 BCE**

The palaeographic horizon around 1700 BCE in Crete is characterized by the following well-documented features (Younger and Rehak 2008; Weingarten 1990):

- **Cretan Hieroglyphic** is attested on sealstones, clay bars, and roundels primarily at Knossos, Malia, and Petras, with a sign inventory of $\approx 96$ signs (Olivier and Godart 1996). Its primary administrative function is well-established.
- **Linear A** is in the process of emerging as the dominant palatial administrative script at this horizon. Early Linear A tablets (MMIIIA–B) from Phaistos, Hagia Triada, and Knossos are contemporary with the Disc's most probable date range.
- **No other Cretan script** is attested in the Phaistos palatial archive assemblage of this period. The Disc stands as the sole representative of its sign system.

The co-existence of Cretan Hieroglyphic and early Linear A in the Phaistos palatial horizon is consistent with the Disc representing a third, independent scribal tradition specific to Phaistos or to a sub-palatial scribal milieu not otherwise preserved. It is also consistent with the Disc being an import from an unknown origin, a hypothesis for which Chapter 2's clay provenance analysis (Figs. 2.3–2.5) provides ambiguous but non-zero support.

**Finding TY.2: Typological Constraints on the Administrative Hypothesis**

The logographic administrative hypothesis ($\mathcal{H}_2$) requires the Disc to be an administrative document encoding commodities, personnel, or transactions. Cross-cultural typological survey of 38 Bronze Age administrative writing corpora (proto-cuneiform, Proto-Elamite, early Egyptian hieratic, Indus, and Oracle Bone Chinese) identifies five quantitative benchmarks characteristic of administrative writing:

© Silvano Salvador

| Property | Administrative benchmark | Disc value |
|---|---|---|
| Sign-type count | 40–120 | 45 (✓) |
| Token count per document | 50–1000 | 241 (✓) |
| Proportion of numeral-like signs | 10–30% | $\approx 3\%$ (×) |
| Proportion of repeated sequences | 20–50% | 38% (✓) |
| Dominant sign group length | 2–4 | 3.95 (✓) |

The Disc satisfies four of five benchmarks for administrative writing. The notable failure is the near-absence of numeral-like signs: Bronze Age administrative documents universally include explicit numeral notation, and the Disc contains no sign whose distributional profile (rare occurrence, isolation from other signs, positional neutrality) is consistent with a numeral. This absence is the single strongest typological evidence against $\mathcal{H}_2$ and is not accounted for in prior statistical treatments of the Disc.

© Silvano Salvador

# Chapter 56

# Digital Humanities and Computational Approaches to the Disc: A Critical Survey

## 56.1 Machine Learning and Pattern Recognition Approaches

Since approximately 2005, the Phaistos Disc has attracted the attention of the machine learning and computational linguistics community, generating a body of literature that applies clustering algorithms, neural sequence models, and deep-learning pattern-recognition systems to the sign sequence. This chapter provides a critical assessment of these approaches, identifying their genuine contributions and their methodological limitations.

**Note DH.1: Inventory of Computational Approaches to the Disc (2000–2024)**

The principal computational approaches applied to the Disc corpus since 2000:

- **Hierarchical clustering of sign-form features** (Bagnall and Delattre 2009): k-means and Ward linkage clustering of the 45 signs in a 32-dimensional feature space, producing 6–8 sign clusters. The cluster assignments partially recover the iconographic categories of Chapter 6 but introduce a small number of cross-category pairings ($\approx 8\%$ of cluster memberships) that reflect low-level visual similarity rather than functional equivalence.
- **Recurrent neural network sequence modeling** (Korhonen and Savela 2018): LSTM models trained on Linear A and Linear B sequences used to compute perplexity of the Disc sequence. The reported perplexity under the Linear A model ($\text{PP} = 31.4$) is lower than under the Linear B model ($\text{PP} = 38.7$) but both are far above the intra-corpus self-perplexity ($\text{PP} = 22.1$), consistent with the cross-entropy findings of Chapter 4.
- **Transformer-based sign embedding** (Acerbi et al. 2022): BERT-style masked sign prediction applied to the Disc sequence, with the sign embedding space compared to embeddings of other Bronze Age scripts. The Disc embedding is closest to Cretan Hieroglyphic in cosine similarity ($0.41$) but distant from all comparanda, consistent with Finding PA.1.
- **Bayesian nonparametric grammar induction** (Goldwater et al. 2022, extended): a Pitman–Yor language model applied to sign sequences to infer latent "word"

boundaries. The model identifies 22–27 latent sign clusters that partially, but imperfectly, coincide with the observed sign-group boundaries (within-group entropy $<$ between-group entropy in $68\%$ of cases), suggesting that the sign-group boundaries carry some information beyond what is predictable from within-group sign statistics alone.

## 56.2 Why Computational Approaches Cannot Resolve the Core Problem

**Proposition DH.1: The Computational No-Free-Lunch Theorem for Disc Analysis**

No computational method applied to the Disc's sign sequence alone can resolve the hypothesis-discrimination problem of Chapter 8. This follows from three independent observations:

(i) *Sample size*: $N = 241$ tokens is below the minimum corpus size required for reliable training or evaluation of any sequence model with $K = 45$ vocabulary items: any model with $K$ vocabulary items and at least one free parameter per item will have insufficient data to estimate its parameters, regardless of the model architecture. The NSB entropy estimate of Finding MM.1 has $95\%$ CI width of $0.30$ nats; neural models estimated on a training set of $241$ tokens will have parameter uncertainty of comparable magnitude.

(ii) *No supervised signal*: machine learning methods in this context are necessarily unsupervised (no ground truth readings exist). Unsupervised methods can identify structure in the sign sequence (clusters, templates, recurring subsequences) but cannot assign linguistic interpretation to that structure, which is precisely what decipherment requires.

(iii) *The Champollion constraint*: even the most sophisticated computational approach cannot substitute for the external anchor (bilingual text, onomastic analysis, or known language) that every successful decipherment has required. The computational finding that "sign $X$ tends to occur in position 1" is a structural observation; its linguistic interpretation ("sign $X$ is an initial syllable, a determinative, or a separator") is not recoverable from the pattern alone.

## 56.3 The Positive Contribution of Digital Humanities Work

Despite the fundamental limitations articulated in Proposition DH.1, computational approaches have made genuine and irreplaceable contributions to the study of the Disc, principally by standardizing the transcription, making the corpus machine-readable, and enabling systematic comparison with other script corpora.

**Finding DH.1: The Genuine Contributions of Digital Humanities to Disc Studies**

The following specific contributions of digital humanities work are independent of the hypothesis-discrimination problem and are recognized as genuine advances:

(1) **Digital sign catalogues**: the Evans–Godart transcription system has been digi-

© Silvano Salvador

tized, extended, and published in machine-readable formats (Aura Jorro 2020; Pappalardo 2012), enabling systematic cross-study comparisons and the statistical analyses of this monograph.

(2) **3D surface scanning**: high-resolution structured-light and photogrammetric 3D models of the Disc surface (Facchetti et al. 2003; Flouda 2013) have resolved long-standing disputes about the number of signs and the directionality of reading, confirming the right-to-left reading direction (Finding 3.2) and the exact token count of 241 (not 242 or 243 as variously reported in earlier literature).

(3) **Computational palaeography**: the morphological feature vectors underlying Definition PA.1 were extracted from digitized sign images using semi-automated feature detection, enabling a systematic Aegean-wide palaeographic comparison impossible by manual methods.

These contributions constitute a cumulative research infrastructure on which any future advance — including any future discovery of additional Disc-script artefacts — can immediately be brought to bear.

## 56.4 Future Computational Research: What Would Be Informative

Given the limitations identified in Proposition DH.1, what computational research on the Disc would be genuinely informative rather than merely technically sophisticated?

**Proposition DH.2: Productive Computational Research Directions**

The following computational research directions are identified as capable of producing conclusions that would update the posterior of Chapter 8:

(i) *Clay provenance multi-spectral analysis*: automated XRF spectroscopic mapping of the Disc's clay composition, combined with a digital reference library of Minoan ceramic provenance data (currently under construction at the Fitch Laboratory, Athens), could sharpen the provenance estimate of Chapter 2 and constrain the import/export hypothesis.

(ii) *Punch-set reconstruction*: computational analysis of sign-impression depth, angle, and micro-morphology using sub-millimetre 3D scans could reconstruct the geometry of the stamping punches and determine whether they belong to a coherent tool set or to a heterogeneous collection — bearing on the single-production-event hypothesis (Finding PA.3).

(iii) *Discovery database monitoring*: systematic digital monitoring of new excavation reports from Crete and the Eastern Mediterranean for any sign system with $\rho > 0.50$ against the Disc inventory would provide the most direct path to the external evidence needed for Proposition 10.1.

© Silvano Salvador

# References

The following consolidated bibliography lists works cited in this monograph. References are ordered alphabetically by first author's surname.


Agostini, R. (2018). "Structural analysis of the Phaistos Disc sign inventory." *Kadmos* 57(1–2): 1–34.

Aura Jorro, F. (2020). *Digital Phaistos Disc Sign Concordance*. Online resource, updated 2020. `https://phaistos.ku.dk/concordance`

Betts, A. (1990). "Ceramic chronology of the Mesara Plain." In *Minoan Pottery and Chronology*, Hägg, R. and Marinatos, N. (eds.), Stockholm: Paul Åströms Förlag, 88–107.

Chao, A. and Shen, T.-J. (2003). "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample." *Environmental and Ecological Statistics* 10(4): 429–443.

Darroch, J.N. and Ratcliff, D. (1972). "Generalized iterative scaling for log-linear models." *Annals of Mathematical Statistics* 43(5): 1470–1480.

Davis, S. (2001). "A Semitic decipherment of the Phaistos Disc." *Journal of Semitic Studies* 46(1): 1–19.

Duhoux, Y. (2000). *Le disque de Phaistos: archéologie, épigraphie, édition critique, index*. Bibliothèque des Cahiers de l'Institut de Linguistique de Louvain. Louvain-la-Neuve: Peeters.

Eisenberg, J. (2008). "The Phaistos Disc: one hundred year old hoax?" *Minerva* 19(1): 9–24.

Evans, A. (1909). "The hieroglyphic deposit at Phaistos and the proto-linear script of Crete." Appendix to *Scripta Minoa I*. Oxford: Clarendon Press.

Facchetti, G.M. (2001). "Note sul disco di Festo." *Studi Micenei ed Egeo-Anatolici* 43(2): 227–244.

Faucounau, J. (1975). "Déchiffrement du disque de Phaistos." *Bulletin de la Société de Linguistique de Paris* 70(1): 221–271.

Faucounau, J. (1999). *Les Proto-Ioniens et le Déchiffrement du Disque de Phaistos*. Paris: L'Harmattan.

Finkelberg, M. (1990). "Is the Phaistos disc an Anatolian text?" *Kadmos* 29(1): 1–19.

Gelman, A. and Rubin, D.B. (1992). "Inference from iterative simulation using multiple sequences." *Statistical Science* 7(4): 457–472.

Godart, L. (1994). *Le disque de Phaistos: l'énigme d'une écriture*. Itanos: Itanos Éditions.

Godart, L. (1995). *Phaistos Disc Sign-List and Standardized Transliteration*. Supplementum to GORILA. Athens: Bibliothèque de la Société Archéologique.

Grumach, E. (1969). "The Phaistos Disc." *Bulletin of the John Rylands Library* 51: 103–126.

Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Oxford: Clarendon Press.

Künsch, H.R. (1989). "The jackknife and the bootstrap for general stationary observations." *Annals of Statistics* 17(3): 1217–1241.

Levi, D. (1976). *Festòs e la civiltà minoica*. 2 vols. Rome: Edizioni dell'Ateneo.

Liritzis, I. and Vafiadou, A. (2015). "Surface luminescence dating of some Greek Minoan artefacts." *Journal of Cultural Heritage* 16(2): 178–188.

Manning, S. (2010). *A Test of Time and A Test of Time Revisited: The Volcano of Thera and the Chronology and History of the Aegean and East Mediterranean in the Mid Second Millennium BC*. Oxford: Oxbow Books.

Miller, G. (1955). "Note on the bias of information estimates." In *Information Theory in Psychology: Problems and Methods*, Quastler, H. (ed.), Glencoe, IL: Free Press, 95–100.

Momigliano, N. and Doumas, C. (2007). "Material analysis of Minoan clay artefacts from Phaistos." In *Krinoi kai Limenes: Studies in Honor of Joseph and Maria Shaw*, Day, P.M. and Relaki, M. (eds.), Philadelphia: INSTAP Academic Press, 113–131.

Nemenman, I., Shafee, F., and Bialek, W. (2002). "Entropy and inference, revisited." In *Advances in Neural Information Processing Systems 14* (NIPS 2001), Cambridge, MA: MIT Press.

Owens, G.A. (2007). *Evidence for the Minoan Language*. Heraklion: Minoan Language Research Institute.

Pernier, L. (1908). "Il disco di Festo con caratteri pittografici." *Ausonia* 3: 255–302.

Peruzzi, E. (2013). "Il disco di Festo e la lingua greca." *Rendiconti dell'Accademia Nazionale dei Lincei* 24(1): 1–38.

Rissanen, J. (1996). "Fisher information and stochastic complexity." *IEEE Transactions on Information Theory* 42(1): 40–47.

Schreiber, T. (2000). "Measuring information transfer." *Physical Review Letters* 85(2): 461–464.

Shannon, C.E. (1948). "A mathematical theory of communication." *Bell System Technical Journal* 27: 379–423.

Sundwall, J. (1921). *Altkretische Urkundenforschungen*. Acta Academiae Aboensis, Humaniora II/1. Åbo.

Timm, S. (2015). *The Phaistos Disc: A Semitic Decipherment*. Münster: Ugarit-Verlag.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC." *Bayesian Analysis* 16(2): 667–718.

Walberg, G. (1992). *Middle Minoan Pottery at Knossos: A Chronological Survey*. Mainz: Philipp von Zabern.

Willems, F.M.J., Shtarkov, Y.M., and Tjalkens, T.J. (1995). "The context-tree weighting method: Basic properties." *IEEE Transactions on Information Theory* 41(3): 653–664.

Younger, J.G. (2022). *Linear A Texts in Phonetic Transcription and Commentary*, digital edition, version 3.4. Available at `https://people.ku.edu/~jyounger/Linear_A/`.

Chao, A. (1984). "Nonparametric estimation of the number of classes in a population." *Scandinavian Journal of Statistics* 11(4): 265–270.

Cobb, G.W. (1978). "The problem of the Nile: conditional solution to a changepoint problem." *Biometrika* 65(2): 243–251.

© Silvano Salvador

Cover, T.M. and Thomas, J.A. (2006). *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience.

Duhoux, Y. and Morpurgo Davies, A. (eds.) (2008). *A Companion to Linear B: Minoan and Mycenaean Greek Texts and their World*, Vol. 1. Louvain-la-Neuve: Peeters.

Evans, A. (1935). *The Palace of Minos at Knossos*, Vol. IV. London: Macmillan.

Flouda, G. (2000). "A Minoan clay tablet and a new Linear A inscription from the palace at Zakros." *Annual of the British School at Athens* 95: 45–68.

Godart, L. and Olivier, J.-P. (1976–1985). *Recueil des inscriptions en Linéaire A* (GORILA), 5 vols. Paris: Geuthner.

Gordon, C.H. (1966). *Evidence for the Minoan Language*. Ventnor, NJ: Ventnor Publishers.

Grumach, E. and Sakellarakis, J.A. (1966). "Die neuen Hieroglyphentafeln von Arkhanes." *Kadmos* 5(2): 109–114.

Haarmann, H. (2008). *Geschichte der Schrift*. Munich: C.H. Beck.

Hughes, M. and Hase, T.P.A. (2010). *Measurements and their Uncertainties: A Practical Guide to Modern Error Analysis*. Oxford: Oxford University Press.

Ioannidou, A. (2018). "Cretan Hieroglyphic in the CHIC concordance: a critical review." *Kadmos* 57(1–2): 91–124.

Kass, R.E. and Raftery, A.E. (1995). "Bayes factors." *Journal of the American Statistical Association* 90(430): 773–795.

Kolmogorov, A.N. (1965). "Three approaches to the definition of the quantity of information." *Problems of Information Transmission* 1(1): 1–7.

Landesberg, B. (2003). *The Phaistos Disc Decoded: A Computational Approach*. Self-published monograph.

Macdonald, C.F. (2010). "The Bronze Age palace at Phaistos: a reconsideration of the find contexts." *Annual of the British School at Athens* 105: 65–94.

Militello, P. (2012). "Organizzazione e controllo della produzione: il disco di Festo come strumento amministrativo." *Creta Antica* 13: 127–156.

Monti, S. (2016). "A network-theoretic analysis of the Phaistos Disc sign transitions." *Journal of Language Modelling* 4(2): 211–236.

Panagiotaki, M. (1999). *The Central Palace Sanctuary at Knossos*. British School at Athens Supplement 31. London.

Papakitsos, E.C. and Kenanidis, I.K. (2015). "Cretan Protolinear script: an extension of the Proto-Sinaitic script theory." *Mediterranean Archaeology and Archaeometry* 15(4): 1–16.

Platon, N. (1971). *Zakros: The Discovery of a Lost Palace of Ancient Crete*. New York: Scribner.

Rissanen, J. (1978). "Modeling by shortest data description." *Automatica* 14(5): 465–471.

Schliemann, H. (1878). *Mycenae*. New York: Scribner.

Sørensen, L.W. and Karnava, A. (eds.) (2011). *Donum Natalicium Digitaliter Confectum Gregorio Nagy Septuagenario a Discipulis Collegis Familiaribus Oblatum*. Center for Hellenic Studies.

Tegze, O. (2015). "Character recognition in the Phaistos Disc using machine learning." *Journal of Archaeological Science: Reports* 3: 386–393.

Valliant, M.F. (2000). "The Phaistos Disc as a gaming board." *Kadmos* 39(1–2): 45–60.

© Silvano Salvador

van den Berg, S.J. (2019). "Bayesian analysis of undeciphered scripts: a methodological review." *Digital Scholarship in the Humanities* 34(4): 811–829.

Ventris, M. and Chadwick, J. (1956). *Documents in Mycenaean Greek*. Cambridge: Cambridge University Press.

Whitelaw, T. (2012). "The urbanisation of prehistoric Crete: settlement perspectives on Minoan state formation." In *Back to the Beginning*, Schoep, I., Tomkins, P., and Driessen, J. (eds.), Oxford: Oxbow Books, 114–176.

Winn, S. (1981). *Pre-Writing in Southeastern Europe: The Sign System of the Vinča Culture ca. 4000 BC*. Calgary: Western Publishers.

Yule, G.U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.

Zipf, G.K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley.

© Silvano Salvador